New Data-Driven Approaches to Text Simplification

Sanja Štajner BSc, MA

A thesis submitted in partial fulfilment of the requirements of the University of Wolverhampton for the degree of Doctor of Philosophy

January 2015

Signature: . . . . . .

Date: . . . . . .

"Knowledge is a process of piling up facts; wisdom lies in their simplification."

Martin H. Fischer

ABSTRACT

Many texts we encounter in our everyday lives are lexically and syntactically very complex. This makes them difficult to understand for people with intellectual or reading impairments, and difficult for various natural language processing systems to process. This motivated the need for text simplification (TS) which transforms texts into their simpler variants. Given that this is still a relatively new research area, many challenges are still remaining. The focus of this thesis is on better understanding the current problems in automatic text simplification (ATS) and proposing new data-driven approaches to solving them.

We propose methods for learning sentence splitting and deletion decisions, built upon parallel corpora of original and manually simplified Spanish texts, which outperform the existing similar systems. Our experiments in adaptation of those methods to different text genres and target populations report promising results, thus offering one possible solution for dealing with the scarcity of parallel corpora for text simplification aimed at specific target populations, which is currently one of the main issues in ATS.

The results of our extensive analysis of the phrase-based statistical machine translation (PB-SMT) approach to ATS reject the widespread assumption that the success of that approach largely depends on the size of the training and development datasets. They indicate more influential factors for the success of the PB-SMT approach to ATS, and reveal some important differences between cross-lingual MT and the monolingual

MT used in ATS.

Our event-based system for simplifying news stories in English (EventSimplify) overcomes some of the main problems in ATS. It does not require a large number of handcrafted simplification rules nor parallel data, and it performs significant content reduction. The automatic and human evaluations conducted show that it produces grammatical text and increases readability, preserving and simplifying relevant content and reducing irrelevant content.

Finally, this thesis addresses another important issue in TS which is how to automatically evaluate the performance of TS systems given that access to the target users might be difficult. Our experiments indicate that existing readability metrics can successfully be used for this task when enriched with human evaluation of grammaticality and preservation of meaning.

CONTENTS

List of Tables

List of Figures

List of Acronyms

ASD	Autism Spectrum Disorders
ATS	Automatic Text Simplification
EAS	Easy Access Sentences
EW	English Wikipedia
FKGL	Flesch-Kincaid Grade Level
ID	Intellectual Disabilities
IR	Information Relevance
LM	Language Model
LWLM	Latent Words Language Model
MID	Mild Intellectual Disabilities
MT	Machine Translation
NLP	Natural Language Processing
PB-SMT	Phrase-Based Statistical Machine Translation
POS	Part of Speech

SEW	Simple English Wikipedia
SMT	Statistical Machine Translation
SRL	Semantic Role Labeling
SVM	Support Vector Machines
TM	Translation Model
TS	Text Simplification
TSM	Tree-based Simplification Model
WAI	Web Accessibility Initiative
WCAG	Web Content Accessibility Guidelines
WSD	Word Sense Disambiguation

ACKNOWLEDGEMENTS

The work on this thesis was an interesting journey. Along the way, there have been many people who enriched it in various ways, either by providing their scientific help, brainstorming discussions, technical and administrative help, or simply by always being there for me, giving me their unconditional love and support. In order to be consistent with the topic of the thesis, I will try to make this section brief and simple, and to only provide the most relevant information.

First of all, my deepest gratitude goes to Prof. Ruslan Mitkov, my director of studies, who introduced me to the world of natural language processing and computational linguistics five years ago, during my Masters studies. In spite of your numerous commitments and academic duties, you always found time to provide me with your support and advice. You made me a better scientist, enabling me to be involved in various aspects of the scientific world, through projects, journals, conferences and summer schools. I will always be grateful for that.

My gratitude also goes to Dr. Constantin Orăsan, my supervisor, who has given me his academic, technical, and administrative support during these years spent at the University of Wolverhampton.

A massive 'thank you' is reserved for Prof. Horacio Saggion, who joined my supervisory team a bit later, but still had a crucial impact on this thesis. In spite of the geographical distance, you were always there for me, supporting me, providing your

feedback, teaching me so much about text simplification and most importantly, believing in my work even in times when I was starting to doubt it. For all of that, and much more, thank you!

The next person to whom I owe my most sincere gratitude is Emma Franklin, my colleague and friend, who not only proof read this whole thesis and most of my papers, but also improved my academic English so much during these years and brought me closer to the English culture. Thank you for always being there for me.

I also wish to express my gratitude to Carmen de Vos Martin and Hannah Béchara who proof read early versions of some of these chapters, and to all those lovely people – Erin Stokes, Alison Carminke, Stephanie Kyle, Iain Mansell, Helen Williams, and Katherine Shepherd – who helped me through the numerous administrative issues I had during my studies.

To all former and current members of the group, my colleagues and friends, thank you for taking part in this journey and for making Wolverhampton a sunnier place. A special 'thank you' goes to Natalia Konstantinova and Victoria Yaneva who taught me some valuable lessons about life, and who always managed to make me feel better and find the right perspective, even in the most difficult moments.

Finally, the biggest 'thank you' is to my parents who gave me the perfect basis for natural language processing and computational linguistics. You planted my love for mathematics, computer science, and linguistics, and I finally managed to combine them all.

CHAPTER 1

INTRODUCTION

Many texts we come across in our everyday lives are lexically and syntactically very complex. This makes them difficult to understand for non-native speakers (Petersen and Ostendorf, 2007; Aluísio et al., 2008), children (De Belder and Moens, 2010), people with intellectual disabilities (Feng, 2009; Saggion et al., 2011), and language-impaired people such as autistic (Martos et al., 2012), aphasic (Carroll et al., 1998; Devlin, 1999), dyslexic (Rello, 2012) and congenitally deaf people (Inui et al., 2003). At the same time, such texts pose obstacles for various natural language processing (NLP) tasks, such as parsing (Chandrasekar et al., 1996), semantic role labelling (Vickrey and Koller, 2008), information retrieval (Beigman Klebanov et al., 2004), and information extraction (Evans, 2011). The benefits of transforming complex texts into their lexically and syntactically simpler variants would thus be two-fold; it would make texts more accessible to wider audiences and improve the performance of various NLP systems.

Since the 1990s, there have been various initiatives which proposed guidelines for making easy-to-read texts (PlainLanguage, 2011; Freyhoff et al., 1998; Mencap, 2002). However, simplification of the existing written material by human editors is both very expensive and time consuming, especially in the case of news articles which are constantly being generated. Therefore, many attempts have been made to completely or

at least partially automate this process. Automatic text simplification (ATS) systems have been proposed for English (Siddharthan, 2006; De Belder and Moens, 2010; Zhu et al., 2010; Woodsend and Lapata, 2011a; Coster and Kauchak, 2011a; Wubben et al., 2012), Spanish (Saggion et al., 2011; Drndarević et al., 2013), and Portuguese (Aluísio and Gasperin, 2010), with recent attempts at Basque (Aranzabe et al., 2012), Swedish (Rybing et al., 2010), Dutch (Ruiter et al., 2010), Italian (Barlacchi and Tonelli, 2013), and French (Brouwers et al., 2014).

Given that automatic text simplification is still a relatively new research area, many challenges are still remaining. The focus of this thesis is on identifying and better understanding the main problems in automatic text simplification and proposing new data-driven approaches to addressing them. Depending on the existing resources, the experiments were performed for English, or Spanish, or both languages.

1.1 Main Problems in Automatic Text Simplification

Our extensive literature review presented in Chapters 2 and 3 identified four main problems in the current state-of-the-art ATS systems:

1. **Parallel corpora for text simplification aimed at specific target populations are very scarce and limited in their size.** Their compilation is very expensive and time-consuming as manual simplification needs to be performed by trained human editors aware of the specific needs of the target population.

2. **Automatic text simplification systems require either a large number of hand-crafted simplification rules or large amounts of parallel data.** The first ATS

systems were rule-based. Data-driven approaches to ATS took the leading role only recently, mostly due to the emergence of Simple English Wikipedia (SEW) which together with the English Wikipedia (EW) offered a large amount of parallel training data (EW-SEW corpus). However, such a large amount of parallel data for text simplification does not exist in any other language. Therefore, rule-based approaches to ATS are still dominant for all languages except English. The main shortcomings of rule-based ATS systems are that they require handcrafting of a great number of simplification rules which are domain- and language-specific, and that they usually have a very limited coverage of lexical simplification rules (De Belder and Moens, 2010). Such systems cannot be easily adapted to different domains and languages.

3. **The existing automatic text simplification systems do not perform sufficient content reduction.** The importance of content reduction in text simplification has been emphasised in several studies (Bautista et al., 2011; Saggion et al., 2011). This is particularly important in the context of ATS aimed at people with intellectual disabilities as they have problems with the memory load required and cannot process large amounts of information (Morgan and Moni, 2008; Gómez, 2011). Some of the recently proposed data-driven ATS systems perform a certain level of content reduction (Coster and Kauchak, 2011a; Zhu et al., 2010; Woodsend and Lapata, 2011a). However, the content reduction achieved by those systems is limited to deleting just a few short sentence parts. It is also not semantically motivated, thus often being erroneous and leading to a change in meaning or the

loss of some relevant information (Narayan and Gardent, 2014). The work of Drndarević and Saggion (2012) confirmed that content reduction is not a trivial task even if it is addressed with a specifically dedicated module. Therefore, we give particular attention to this problem.

4. **There is no well-established methodology for evaluating text simplification systems and comparing their performance.** With the emergence of automatic text simplification systems, the question we are faced with is how to automatically evaluate their performance given that access to the target users might be difficult. Feng et al. (2009), Petersen and Ostendorf (2009), and Schwarm and Ostendorf (2005) raised some doubts over the suitability of using the absolute value of readability indices as a measure of user comprehension of simplified texts. The authors showed that some cognitively motivated features (e.g. entity mentions, lexical chains, etc.) are better correlated with comprehension of texts by people with mild intellectual disabilities. Nevertheless, the absolute value of readability indices is often used in automatic evaluation of ATS systems, e.g. those systems proposed by Woodsend and Lapata (2011a) and by Zhu et al. (2010).

1.2 Research Questions

Based on the main problems in automatic text simplification, identified in the previous section, we pose four research questions (RQ):

- **RQ 1:** Is it possible to adapt an already existing TS system aimed at a specific target audience to a TS system aimed at a different target population?

It is generally agreed that there are more factors which unify the needs of different target groups than those which separate them (Nomura et al., 1997). In spite of this, there have been no studies investigating whether it is possible to adapt an already existing ATS system aimed at a specific target audience in such a way that it can perform text simplification necessary for another target population. Chapter 4 presents the first steps in searching for that answer, focusing on decision-making systems for sentence splitting and sentence deletion in text simplification systems for Spanish. Chapter 7 also contributes to answering this question by offering easy ways to compare complexity reduction achieved by various TS systems and the complexity reduction necessary when making texts more accessible to specific target populations.

- **RQ 2:** Is it possible to build a TS system which does not require large amounts of parallel data or handcrafted rules, but rather exploits some already existing NLP tools and can easily be adapted to different languages?

  Chapter 5 presents the first steps in answering this question, exploring the use of phrase-based statistical machine translation (PB-SMT) models in text simplification. Previous studies showed that such models can be used successfully for this task. However, there is a widespread assumption that the PB-SMT approach requires large training and development datasets in order to be successful. We investigated whether the success of that approach mostly depends on the sizes of the datasets or there might be some other more important factors which would allow such systems to be successful even when trained on smaller datasets.

Chapter 6 explores this research question from another angle, proposing a system for simplifying news stories in English. The system is built upon the state-of-the-art event extraction system (Glavaš and Šnajder, 2014). In its current state, the proposed ATS system performs syntactic simplification with significant content reduction. It produces separate sentences for each event mention, at the same time erasing all sentence parts, and entire sentences, which do not belong to any factual events. Our ATS system does not require any parallel corpora, and it can be easily adapted to different languages and domains under the condition that the adequate event extraction systems exist.

- **RQ 3:** Is it possible to build a TS system which, in addition to simplifying the given text, also performs a significant content reduction by deleting irrelevant information?

  We approach this question from two different angles. Chapter 4 proposes a corpus-based decision-making system for detecting sentences which should be deleted during simplification. This system can be added to some of the already existing rule-based TS systems for Spanish, e.g. the system built under the Simplext project (Saggion et al., 2011). Chapter 6 proposes a semantically-motivated, event-based text simplification system for English which performs significant content reduction.

- **RQ 4:** Could some of the already existing readability indices be used for the automatic evaluation of text simplification systems?

  In Chapter 7, we show that there is a significant correlation between readability

indices and the linguistically motivated features, and suggest the use of relative values of readability indices for automatic evaluation of text simplification systems.

1.3 Contributions

This thesis makes a number of novel contributions to text simplification by critically analysing the existing approaches, and proposing new ATS systems and new evaluation methods. In this section, we present only the main contributions (**C1 – C6**), while more detailed lists of contributions can be found in the summary sections of each chapter.

C1: A comprehensive overview of linguistic obstacles to human comprehension and machine processing from various aspects: psycholinguistics, existing guidelines for producing easy-to-read texts, and data-driven approaches based on the analysis of parallel corpora (original texts and their manual simplifications).

C2: A critical overview of existing ATS systems and identification of their main shortcomings, which motivate our research questions.

C3: We propose a new feature set which leads to the state-of-the-art performance of two decision-making modules in ATS systems for Spanish: (1) classification of original sentences into those to be deleted and those to be kept during simplification; and (2) classification of original sentences into those to be split and those to be left unsplit during simplification. The main potential of these classification systems lies in enriching the state-of-the-art rule-based text simplification systems (such as the ATS system for Spanish proposed under the Simplext project, for example) if they are included at the beginning of the simplification pipeline. The proposed classification systems can

eliminate unnecessary sentences (thus introducing a content reduction module which is currently not present in any of the rule-based systems) and detect the sentences which need to be split (and thus send them to a dedicated syntactic simplification module). Additionally, we show that:

- Sentence deletion decisions trained on one type of TS corpus cannot be successfully applied to different text genres and ATS aimed at a different target population.

- Sentence splitting decisions trained on one type of TS corpus can be successfully applied to different text genres and ATS aimed at a different target population.

C4: An extensive investigation of the phrase-based statistical machine translation (PB-SMT) approach to TS which indicated the following:

- The BLEU score (Papineni et al., 2002) is not a good measure of the performance of a standard PB-SMT model in TS, as it mainly reflects the similarity between the original sentences and their simplified versions in the test set and not the actual system's performance (due to the important differences between cross-lingual MT and the monolingual MT used in TS).

- The type of the training and development datasets (parallel or comparable corpora) does not have any impact on the success of a standard PB-SMT model in text simplification.

- The size of the training and development datasets does not significantly influence the performance of a standard PB-SMT model in TS (in general).

- The similarity of the original sentences and their simplified versions (in terms of sentence-wise BLEU score) in the training and development datasets significantly influences the quality of the output generated by a PB-SMT system in all three aspects (grammaticality, meaning preservation, and simplicity).

C5: We propose a new text simplification system (EventSimplify) which simultaneously simplifies and reduces the content of a given text. The system does not require any parallel TS data nor large numbers of handcrafted simplification rules. It is semantically motivated and built upon a state-of-the-art event extraction system. The performance of the system is comparable to the state-of-the-art ATS systems in English (which require large parallel TS datasets), and it can be easily adapted to any language under the condition that there is a robust enough event extraction system for that language.

C6: We show that some of the already existing readability indices have a good correlation with the possible obstacles to reading comprehension and thus could be used for the automatic evaluation of simplicity achieved by text simplification systems. We suggest the use of readability indices in text simplification for an easy comparison of:

1. Original and simplified texts in order to assess either the necessary complexity reduction (if comparing original texts with the manually simplified ones) or the achieved complexity reduction (if comparing original texts with the automatically simplified ones);

2. Different text simplification systems (i.e. the level of simplification achieved by different TS systems);

3. Automatically simplified texts with the manually simplified ones (in order to as-

sess whether the automatic simplification achieves the same level of simplification as the manual one);

4. Manually simplified texts with a 'gold standard' (easy-to-read texts which were originally written with the target population in mind) with the aim of assessing whether the manually simplified texts reach the simplicity of the 'gold standard', and thus comply with the easy-to-read standards.

1.4 Outline of the Thesis

Chapter 2 introduces the need for TS systems, presenting some of the obstacles to human comprehension (Section 2.1) and to machine processing (Section 2.2). It compares existing guidelines for writing easy-to-read texts which would be more accessible to people with various reading or intellectual impairments. This chapter also identifies possible problems in following easy-to-read guidelines, thus calling into question the reliability of simplified texts obtained in this manner (Section 2.3). Section 2.4 approaches the detection of necessary transformations in text simplification from a data-driven perspective, analysing parallel corpora of original and manually simplified texts aimed at various target populations. This chapter introduces the first original contribution (**C1**).

Chapter 3 presents different approaches to text simplification used in previous studies. It compares various ATS systems and draws attention to their main strengths and weaknesses, opening new research avenues to be explored in this thesis (**C2**).

Chapter 4 reports on two sets of experiments: (1) classification of original sentences into those to be *deleted* and those to be *kept*; and (2) classification of original sentences

into those to be *split* and those to be left *unsplit*. The proposed set of features and classification algorithms outperforms previously proposed similar systems. This chapter brings several important insights to decision-making modules in text simplification systems (**C3**), and addresses two research questions (**RQ 1** and **RQ 2**).

Chapter 5 presents several sets of experiments which lead to a better understanding of a PB-SMT approach to text simplification (**C4**), addressing the second research question (**RQ2**). Based on the experiments in three languages (English, Spanish, and Brazilian Portuguese), we reject the widespread assumption that the success of a PB-SMT approach largely depends on the size of the training and development datasets, and indicate the more probable causes of the success of such a PB-SMT approach to TS reported in previous studies. In this chapter, we also show how the sentence pairs in the training and development datasets can be filtered to improve the 'translation' performance, and we reveal some important differences between cross-lingual MT and the monolingual MT used in TS.

Chapter 6 proposes a new text simplification system (EventSimplify) which simultaneously reduces and simplifies the content of a given text in English (**C5**). The system employs a semantically motivated, event-based simplification approach built upon a state-of-the-art event extraction system. In its current state, the system performs syntactic simplification and significant content reduction. In future, the system can be enriched with a lexical simplification module or combined with the most successful PB-SMT module proposed in the previous chapter. Chapter 6 addresses two research questions (**RQ 2** and **RQ 3**).

Chapter 7 investigates whether some of the already existing readability formulae

have a good correlation with the possible obstacles to reading comprehension and thus could be used for the automatic evaluation of simplicity achieved by text simplification systems (**C6**). It reports on experiments in English and Spanish, reporting comparable results in both languages. In Sections 7.4 and 7.5, we show that there is a significant correlation between readability indices and the linguistically motivated features we proposed. Based on those findings, in Section 7.6, we suggest several possible uses of readability indices in text simplification. This chapter addresses the last research question (**RQ 4**).

Chapter 8 revisits the main research questions and original contributions of this thesis. It summarises the experiments and main findings of each chapter, comments on their potential impact on future text simplification studies, and proposes new research avenues.

CHAPTER 2

DETECTION OF NECESSARY TRANSFORMATIONS FOR TEXT SIMPLIFICATION

In order to find the best strategy for an automatic text simplification system, one should first understand the possible linguistic obstacles which need to be removed in order to make the text easier for humans to comprehend and easier for machines to process. Therefore, in this chapter, we briefly introduce some of the linguistic obstacles to human comprehension (Section 2.1) and to machine processing (Section 2.2), and the existing guidelines for accessible writing (Section 2.3). We also present previous studies which try to detect the necessary transformations for automatic text simplification by analysing the existing corpora of original and manually simplified texts for various target populations (Section 2.4).

2.1 Linguistic Obstacles to Human Comprehension

Access to written information for people with intellectual impairments and people with various reading and comprehension difficulties is a fundamental human right, which enables them to have better inclusion into society. This was stated by the Convention on the Rights of Persons with Disabilities, adopted by the United Nations in 2006. However, the vast majority of texts we come across in our everyday lives are syntactically and lexically too complex and can be difficult to understand by non-native speakers (Pe-

tersen and Ostendorf, 2007; Aluísio et al., 2008), children (De Belder and Moens, 2010), people with intellectual disabilities (Feng, 2009; Saggion et al., 2011), and language-impaired people such as autistic (Štajner et al., 2012; Martos et al., 2012), aphasic (Carroll et al., 1998; Devlin, 1999), dyslexic (Rello, 2012) or congenitally deaf people (Inui et al., 2003).

For example, long sentences, noun compounds and long sequences of adjectives, e.g. "twenty-five-year-old blond-haired mother-of-two Jane Smith" (Carroll et al., 1998), which are some genre-specific characteristics of newswire texts, can cause problems for people with aphasia[1] (Carroll et al., 1998), autism spectrum disorders – ASD[2] (Martos et al., 2012), and intellectual disabilities (Feng, 2009).

Some particular sentence constructions, such as syntactic constructions which do not follow the canonical subject-verb-object structure (e.g. passive constructions) may also be an obstacle for people with aphasia (Devlin, 1999), or ASD (Martos et al., 2012). These are frequently used in newswire texts in order to present the information in a more sensational way. For example, it is more common to find a sentence in passive *"A bid to build an incinerator on local wasteland was today accepted by the council"* instead of the more straightforward version *"The council today accepted a bid to build an incinerator on local wasteland"* (Carroll et al., 1998), which would be more easily understood by both aphasic and autistic people. Even more difficult for aphasic people can be those sentences which are semantically reversible, e.g. *"The boy was kissed by*

[1] Aphasia is a language processing disability usually caused by a stroke or a head injury. The language impairments of people with aphasia are quite diverse, but many aphasic people are very likely to encounter problems in understanding written text at some point (Carroll et al., 1998).

[2] ASD is a set of neurodevelopmental disorders characterised by qualitative impairment in communication and stereotyped repetitive behaviour. People with ASD have deficits in the comprehension of speech and writing. (Štajner et al., 2012)

the girl" (Carroll et al., 1999).

Infrequent words make the text difficult to comprehend for people with aphasia (Devlin, 1999), and ASD (Norbury, 2005; Martos et al., 2012). Use of more frequent words does not improve comprehension but reduces the reading time in people with dyslexia (Rello et al., 2013). When it comes to students with intellectual disabilities, the existing studies show contradictory findings. Fajardo et al. (2014) found no effects of the word frequency on the comprehension scores (neither literal nor inferential[3]) in students with intellectual disabilities. Karreman et al. (2007) reported both literal and inferential comprehension of the group of people with intellectual disabilities higher in the adapted version than in the non-adapted version of a website (see Section 2.3.1 for more details on this study). One possible reason for those – at first sight contradictory – findings might lie in the fact that the adapted and non-adapted websites used in the study of Karreman et al. (2007) differed in a number of linguistic elements (e.g. length of words and sentences, frequency and abstractness of words, tense of sentences, etc.). This makes it difficult to distinguish which (or which set) of those elements actually caused better comprehension of adapted websites (Fajardo et al., 2014).

At the discourse level, people with autism or intellectual disabilities may also have problems finding the main idea (Martos et al., 2012; Feng, 2009), resolving the anaphors (Martos et al., 2012), inferring information (Martos et al., 2012; Feng, 2009) and comprehending the text in dialogue format (Martos et al., 2012; Drndarević and Saggion, 2012). Additionally, people with intellectual disabilities have problems processing and

[3]Literal comprehension tests the actual meaning of single propositions while the inferential comprehension tests the integration between text segments or between text and prior knowledge. For a more detailed explanation of differences between literal and inferential comprehension and examples see the study by Fajardo et al. (2014).

retaining large amounts of information (Feng, 2009; Fajardo et al., 2014). Several studies have shown that long texts can affect self-efficacy and reading motivation in students with intellectual disability (Morgan and Moni, 2008; Gómez, 2011). The study of Gernsbacher and Faust (1991) indicated that adult poor comprehenders have difficulties in suppressing irrelevant information. Therefore, text simplification systems aimed at those target populations should not only simplify the written content (by using simpler synonyms and splitting long and complex sentences into several simple ones), but should also perform some kind of content reduction (discarding irrelevant information) in order to reduce the memory load necessary for understanding the given text.

2.2 Linguistic Obstacles to Machine Processing

Long and complicated sentences are not only an obstacle for comprehension by humans, but they are also a stumbling block for many NLP systems such as parsing (Chandrasekar et al., 1996), machine translation (Chandrasekar, 1994), information extraction (Beigman Klebanov et al., 2004; Evans, 2011), and semantic role labelling (Vickrey and Koller, 2008). In all those studies, sentence simplification was suggested as a preprocessing step in order to improve the performance of those NLP systems. Chandrasekar et al. (1996) gave the example of a typical sentence in newswire texts (1), and its simplified multi-sentence version obtained by manual simplification (2) to illustrate the potential of using simplified sentences in various NLP systems:

1. *"The embattled Major government survived a crucial vote on coal pits closure as its last-minute concessions curbed the extent of Tory revolt over an issue that generated unusual heat in the House of Commons and brought the miners to London*

streets."

2. *"The embattled Major government survived a crucial vote on coal pits closure. Its last-minute concessions curbed the extent of Tory revolt over the coal-mine issue. This issue generated unusual heat in the House of Commons. It also brought the miners to London streets."*

Chandrasekar et al. (1996) pointed out that simple sentences generate a smaller number of possible parse trees and involve fewer constituents. This results in reduced ambiguity in attachment of constituents and leads to a faster and less ambiguous parsing. These kinds of simpler sentence structures and reduced ambiguity can also lead to improvements in the quality of machine translation systems (Chandrasekar, 1994).

Vickrey and Koller (2008) applied semantic role labelling (SRL) to the output of a rule-based text simplification system which comprises 16 rule categories (sentence normalisation, sentence extraction, passive to active transformation, rewriting of appositions, etc.). They reported a statistically significant 1.2% F-measure improvement over a strong baseline on the CONLL-2005 SRL task[4].

Beigman Klebanov et al. (2004) introduced the notion of Easy Access Sentences (EAS) which are easy to retrieve information from. Each EAS is a grammatical sentence with one tensed verb reporting a piece of information explicitly or implicitly present in the original text, in which pronouns are substituted with the appropriate names.

Evans (2011) showed that the use of simplified sentences (whose simplification is based on detection of the commas, coordinating conjunctions, and adjacent comma-

[4]http://www.lsi.upc.edu/ srlconll/st05/st05.html

conjunction pairs) improves information extraction in medical texts.

2.3 Proposed Guidelines

Since the late nineties, several initiatives have raised awareness of the complexity of the vast majority of written documents and the difficulties they pose to people with any kind of reading or learning impairments. These initiatives proposed various guidelines for writing in a simple and easy-to-read language which would be equally accessible to everyone. An extensively discussed question is how much the needs of different target populations overlap or not (Nomura et al., 1997). It is generally agreed that there are more factors which unify different target groups than those which separate them (Nomura et al., 1997).

In this section the focus will be on the three most explicit guidelines (in terms of verbal content, not the layout): the "Federal Plain Language Guidelines" (PlainLanguage, 2011), "Make it Simple, European Guidelines for the Production of Easy-to-Read Information for people with Learning Disability" (Freyhoff et al., 1998), and "Am I making myself clear? Mencap's guidelines for accessible writing" (Mencap, 2002). Although aimed at different target populations, they all share the same main ideas for accessible writing.

The Plain Language Action and Information Network (PLAIN)[5] developed the first version of the "Federal Plain Language Guidelines" in the mid-90s and have revised it every few years since then. Their original idea was to help writers of governmental documents (primarily regulations) to write in a clear and simple manner so that the users

[5] http://www.plainlanguage.gov/

can: "find what they need, understand what they find; and use what they find to meet their needs." (PlainLanguage, 2011).

"Make it Simple, European Guidelines for the Production of Easy-to-Read Information for people with Learning Disability" was produced by Inclusion Europe[6] in order to assist writers in developing texts, publications and videos that are more accessible to people with intellectual disabilities and other people who cannot read complex texts, and thus enable those people to be better protected from discrimination and social injustice.

"Am I making myself clear? Mencap's guidelines for accessible writing"[7] were produced by the UK's leading organisation working with people with a learning disability[8]. Their goal is to help in editing and writing accessible material for that specific target population.

All of these guidelines are concerned with both verbal content of documents and their layout. As we are interested in text simplification and not in text representation, we will concentrate only on the first part (verbal content of documents). Table 2.1 contains the main rules of the following three guidelines: "Make it Simple" (Freyhoff et al., 1998), "Am I making myself clear?" (Mencap, 2002), and "Federal Plain Language Guidelines" (PlainLanguage, 2011).

For each rule in the first column of the table, the following three columns 'Simple', 'Clear', and 'Plain' contain 'yes' if this rule is present in the corresponding guidelines ("Make it Simple", "Am I making myself clear?" and "Federal Plain Language Guidelines", respectively). Value '(yes)' is used when the rule is not explicitly present in the

[6] http://inclusion-europe.org/

[7] http://november5th.net/resources/Mencap/Making-Myself-Clear.pdf

[8] In easy-to-read guidelines, terms 'intellectual disability' and 'learning disability' are used interchangeably (Nomura et al., 1997)

Table 2.1: Rules for verbal content of documents

Rule	Simple	Clear	Plain
Use active tense (instead of passive)	yes	yes	yes
Use the simplest form of a verb*	(yes)		yes
Avoid hidden verbs (i.e. verbs converted into a noun)			yes
Use 'must' to indicate requirements			yes
Use contractions where appropriate			yes
Don't turn verbs into nouns			yes
Use 'you' to speak directly to readers	yes	yes	yes
Avoid abbreviations	yes		yes
Use short, simple words	yes		yes
Omit unnecessary words			yes
Avoid definitions as much as possible			yes
Use the same term consistently		yes	yes
Avoid legal, foreign and technical jargon	yes	yes	yes
Don't use slashes			yes
Write short sentences	yes	yes	yes
Keep subject, verb and object close together			yes
Avoid double negatives and exceptions to exceptions	(yes)		yes
Place the main idea before exceptions and conditions			yes
Cover only one main idea per sentence	yes	yes	
Use examples (avoid abstract concepts)	yes		yes
Keep the punctuation simple	yes	yes	
Be careful with figures of speech and metaphors	yes		
Use the number and not the word	yes	yes	
Avoid cross references	yes		yes

*Use present tense and not conditional or future

corresponding guidelines, only implicitly. This allows us to have a quick overview of intersecting rules suggested by these guidelines which were intended for slightly different purposes and target audiences.

As can be noted from Table 2.1, all three guidelines share similar instructions for accessible writing, some of them more detailed than others. For example, they all ad-

vise the writer to use the active voice instead of passive, use short, simple words and omit unnecessary words, write short sentences and cover only one main idea per sentence. However, the "Federal Plain Language Guidelines" also specify to use contractions where appropriate, avoid hidden verbs (i.e. verbs converted into a noun), and place the main idea before exceptions and conditions, while the other two guidelines do not go into many details. Some of the instructions, e.g. to use the simplest form of a verb (present and not conditional or future), or to avoid double negatives and exceptions to exceptions are not present in the Mencap's guidelines for accessible writing (column 'Clear' in Table 2.1), while at the same time being implicitly present in the "Make it Simple" guidelines (column 'Simple' in Table 2.1), and explicitly present in the "Federal Plain Language Guidelines" (column 'Plain' in Table 2.1).

The rules for Plain English are the most detailed, providing examples of 'don't say' and 'say' for each of the rules. A few examples of those rules, for avoiding long noun strings and resolving of pronouns, are given in Table 2.2.

Table 2.2: Examples of rules (PlainLanguage, 2011)

Don't say	Instead, say
Underground mine worker safety protection procedures development	Developing procedures to protect the safety of workers in underground mines
Draft laboratory animal rights protection regulations	Draft regulations to protect the rights of laboratory animals
After the Administrator appoints and Assistant Administrator, he or she must ...	After the Administrator appoints an Assistant Administrator, the Assistant Administrator must ...

2.3.1 Validation of the "Make it Simple" Guidelines

The aim of the W3C Web Accessibility Initiative (WAI) guidelines is to make websites more accessible for people with various disabilities (W3C, 2008). They gained the status of formal requirements in many countries (Karreman et al., 2007). A great part of these guidelines is designed to make web information accessible for users with visual and motor disabilities, though the document also claims to be directed at people with reading difficulties or non-native speakers (Karreman et al., 2007). However, these guidelines provide much less information about how to make web content accessible for people with intellectual disabilities which affect language skills (Karreman et al., 2007). One of the highest priority checkpoints (checkpoint 14.1) of the guidelines states that the clearest and simplest language should be used for a site's content: *"Use the clearest and simplest language appropriate for a site's content"* (Web Content Accessibility Guidelines 1.0[9])

Therefore, Karreman et al. (2007) investigated whether the application of the "Make it Simple" guidelines (Freyhoff et al., 1998) to the website's content would enhance its usability for users with intellectual disabilities. Additionally, they investigated whether the application of these guidelines would have a negative effect on users without disabilities, as WAI guidelines state that creation of multiple versions of the same website should be avoided whenever possible. Karreman et al. (2007) prepared two versions of a website in Dutch, the original one and the one adapted according to the "Make it Simple" guidelines. The original website was based on a leaflet written for the care provider organisation, describing its main services and activities in five sections. The adapted

[9] http://www.w3.org/TR/WCAG10/

version was evaluated by two experts (a specialist in care for people with intellectual disabilities and a web communication expert) who assessed whether the easy-to-read guidelines were applied correctly, and by two people with intellectual disabilities. The two versions of the website were further tested for efficiency (searching and reading time) and effectiveness (comprehension) by 40 participants, 20 with diagnosed intellectual disabilities and 20 without. The results demonstrated that the adaptation of the website according to the guidelines enhanced the efficiency and effectiveness for both groups of participants.

2.3.2 Use of Guidelines in Manual Text Simplification

Although it was proved that adaptation of texts following the "Make it Simple" guidelines improves both reading time and comprehension for all readers (Karreman et al., 2007), there has hardly been any work devoted to how those general guidelines can be applied in text simplification (Bautista et al., 2011). We addressed this issue in two studies (Drndarevic et al., 2012; Mitkov and Štajner, 2014), discussing some of the problems in following the guidelines for manual simplification of texts.

We analysed the manual simplifications of 40 news stories in Spanish performed by trained editors following a series of easy-to-read guidelines derived by a group of experts for the purpose of the Simplext project[10] (Drndarevic et al., 2012). The focus was on lexical substitution of reporting verbs (RepV). We found that the original texts contained ten different RepV which were all substituted with the simpler *decir* (say) at least once. The simplified texts contained only three RepV other than *decir* (say):

[10] www.simplext.es

anunciar (announce), *señalar* (point out), and *afirmar* (confirm), all of which were preserved from the original texts.

The repetition of the same word (i.e. *decir*) was avoided in consecutive sentences probably for stylistic purposes (given that the provided guidelines did not specifically address the issue of reporting verbs, but rather only suggested the use of more frequent and shorter words instead of a difficult original word). This is illustrated in the following paragraph from one of the simplified texts (the first sentence is the headline):

> *"El PSOE* **afirma** *que España pierde a un "gigante de la escena" con la muerte de Manuel Alexandre. Muere el actor Manuel Alexandre. El Partido Socialista Obrero Español* **señaló** *su pena por la muerte del actor. El Partido Socialista* **dijo** *que Manuel Alexandre ha sido un extraordinario actor. Ha sido un actor que ha participado en los momentos más importantes del cine español. El Partido Socialista también* **ha dicho** *que el actor amaba su trabajo."* [The SSWP **confirms** that Spain has lost a "giant on stage" with Manuel Alexandre's death. The actor Manuel Alexandre dies. The Spanish Socialist Workers' Party **indicated** their grief at the actor's death. The Socialist Party **said** that Manuel Alexandre was an extraordinary actor. He was an actor that participated in the most important moments of Spanish cinematography. The Socialist Party also **said** that the actor loved his job. (Drndarevic et al., 2012)

Although human editors were not fully consistent in using *decir* instead of any other reporting verb, the automatic text simplification system built under the Simplext project

(Drndarević et al., 2013) substitutes all RepV with *decir* (say). The decision was justified with the fact that *decir* is both the most common and the most general reporting verb (Quirk et al., 1985; Bosque Muñoz and Demonte Barreto, 1999) and shorter than any of its semantic equivalents, which complies with the rules present in the "Make it Simple" guidelines (Freyhoff et al., 1998). The authors also found that substitution of any RepV with *decir* eliminates polysemy, as is the case with the verb *indicar*, which in Spanish means both 'point' (the literal meaning) and 'point out' (non-literal meaning). As stated in WCAG 2.0 guidelines (W3C, 2008), use of non-literal meaning should be avoided in easy-to-read writing.

Particularly interesting was the case in which the verb *anunciar* (announce) was kept (instead of being replaced by the verb *decir* (say)), as a consequence of giving preference to a syntactic simplification over a lexical simplification (Table 2.3).

Table 2.3: An example of giving preference to syntactic simplification (over lexical)

Version	Sentence
Original	*"El Museo del Prado acogerá en 2014 una gran exposición dedicada a El Greco, con motivo del IV centenario del fallecimiento del pintor,* **según anunció** *este martes la presidenta de la Comunidad de Madrid, Esperanza Aguirre."* [The mayor of Madrid, Esperanza Aguirre, **announced** this Tuesday that in 2014 the Prado Museum is going to house a large exhibition dedicated to El Greco, motivated by the fourth centenary of the painter's death.]
Simplified	*"Esperanza Aguirre, presidenta de la Comunidad de Madrid,* **anunció la exposición**." [Esperanza Aguirre, the mayor of Madrid, **announced the exhibition**.]

In this case, if the original verb was substituted with its simpler version, it would result in a syntactically more complex output. This example illustrates one of the main

shortcomings of the existing guidelines, which, to the best of our knowledge, has never been raised before – the lack of hierarchical structure of the rules. None of the guidelines states which rules should be given preference in the case of the editor needing to choose between two of them. This problem is especially pronounced in those guidelines which consist of a great number of rules, such as those used in the Simplext project (which consist of 28 main rules with additional sub-rules for most of the main rules).

We further investigated the influence of the number of rules in the guidelines on the performance of the human editors who produce manual simplifications following those guidelines by contrasting two sets of rules: (1) the Simplext guidelines (taking into consideration only the 28 main rules); and (2) the six rules we proposed (Mitkov and Štajner, 2014). The results indicated that the smaller set of rules leads to a production of equally simplified texts (in terms of readability) as the larger set of rules. At the same time, use of the smaller set of rules results in quicker and more consistent and reliable output. When using the 28 Simplext rules, all annotators were choosing different rules for simplification of the same sentences which led to very different outputs (Table 2.4). The analysis of rules used by each editor (all three editors were requested to list the rules used for adaptation of each sentence during the simplification process), further supported the claim about more consistent and reliable output when the smaller set of rules is used (Mitkov and Štajner, 2014).

These results raise doubts about taking one annotator's output as a 'gold standard' in text simplification, no matter how well the simplification guidelines were developed. Our findings indicate that a parallel TS corpus (of original and manually simplified texts) might not be consistent if the texts were simplified by different editors, or it might

Table 2.4: Differences in the simplification outputs by three annotators

Version	Sentence
Original	*"Esta reforma prevé que todos los delitos relacionados con la seguridad vial (como exceso de velocidad o conducir bajo los efectos del alcohol, las drogas, sin carné o sin puntos) pueden conllevar el decomiso del vehículo, si bien la decisión dependerá del juez."* [This reform will envisage that all crimes related to road safety (such as speeding, driving while under the effects of alcohol or drugs or driving without a licence or points) could result in confiscation of the vehicle, although the decision to do so depends on the judge.]
Annotator 1	*"El cambio del Código Penal dice que la decisión de embargar el coche o moto dependerá del juez."* [The change of the penal code says that the decision to confiscate the car or motorbike depends on the judge.]
Annotator 2	*"Esta reforma prevé que todos los delitos relacionados con la seguridad vial como exceso de velocidad o conducir bajo los efectos del alcohol, las drogas, sin carné o sin puntos. Los delitos pueden conllevar la retirada del vehículo pero la decisión dependerá del juez."* [This reform will envisage that all crimes related to road safety such as speeding or driving under the effects of alcohol, drugs, without a license or points. The crimes could result in confiscation of the vehicle but the decision depends on the judge.]
Annotator 3	*"La reforma del Código Penal prevé que todos los delitos relacionados con la seguridad vial pueden dar lugar a la pérdida del vehículo, aunque la decisión dependerá del juez."* [The penal code reform will envisage that all crimes related to road safety could result in loss of the vehicle, although the decision depends on the judge.]

be biased if all texts were simplified by one person. These results, however, need to be treated with caution given the small number of texts used (10 original texts and their corresponding manual simplifications) and the small number of human editors involved (only three).

2.4 Data-Driven Detection of Necessary Transformations for Automatic Text Simplification

Several existing parallel corpora of original texts and their manual simplifications were used to determine the necessary transformations in TS: for children (Bautista et al.,

2011); for people with intellectual disabilities (Drndarević and Saggion, 2012); for language learners (Petersen and Ostendorf, 2007); for people with low literacy (Gasperin et al., 2009); and for various readers (Coster and Kauchak, 2011b). Those five studies contain the most important work on learning necessary transformations from parallel corpora aimed at specific target populations. Unfortunately, those studies are not directly comparable, either because they focus on different types of transformations, or because they address different languages.

2.4.1 Taxonomy of Transformations

Bautista et al. (2011) used the parallel corpus created by Barzilay and Elhadad (2003) to investigate which kind of simplification transformations need to be applied to an original text in order to obtain its easy-to-read version. The corpus comprises two aligned collections from the Encyclopedia Britannica and Britannica Elementary. The latter is aimed at children and thus contains one-to-two page entries. Although the authors acknowledged that the Britannica Elementary was not created for people with literacy problems which might lead to important differences in the types of transformations observed, they believe that their findings should be considered as a preliminary study. The authors identified the following five types of transformations:

1. Lexical transformations (the use of synonyms, the replacement of words with easy-to-read alternatives).

2. Syntactic transformations that do not affect the semantics.

3. Deletion of non-relevant information.

4. Addition of extra information in the simplified version used to better explain difficult concepts.

5. Complete rewrite of the original sentence (paraphrase).

Furthermore, they created a taxonomy of these transformations (Figure 2.1).

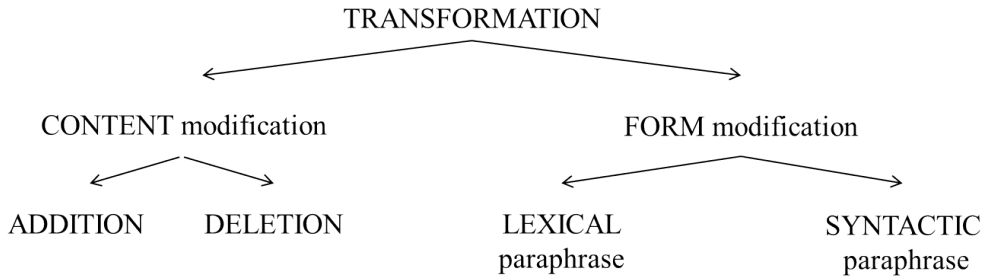


Figure 2.1: Taxonomy of tranformations

The results of the analysis of the 320 aligned sentences between the two versions of Britannica showed that the majority of transformations (51.35%) were deletions, followed by paraphrases (37.95%) and that addition of clarifying information was the least used transformation (10.71%). The total number of detected transformations was 448, leading to the conclusion that more than one transformation is usually applied in each sentence. A more detailed analysis indicated that lexical and syntactic paraphrases were applied in a similar proportion – lexical (52.95%) and syntactic (47.05%). It is interesting to note that our examination of manually simplified Spanish newswire text for people with intellectual disabilities (Štajner et al., 2013) led to similar conclusions, although performed for a different language and different target users. Lexical and syntactic paraphrases were relatively equally present (22% and 25%, respectively), while

the majority of transformations (44%) were applied to lexical and syntactic levels simultaneously. Similar to the work of Bautista et al. (2011), we observed a higher number of transformations (468) than original sentences (247).

Bautista et al. (2011) further reported that among the three observed syntactic paraphrases (passive to active, perfect to simple tense, and transformation of sentence structure), transformations of sentence structure were the most frequent (29.41%). Among the three observed lexical paraphrases (noun, adjective and verb synonyms), noun and verb synonyms were used with a similar frequency (21.17% and 20.59%, respectively) while adjective synonyms were used less frequently (11.18%).

Based on these results, Bautista et al. (2011) proposed an automated simplification system which would rely on:

- Summarisation methodology to delete unnecessary information;
- WordNet (Fellbaum, 2010) as a source of possible lexical paraphrases;
- A rule-based transformation of parse trees obtained automatically using the Stanford Parser (Klein and Manning, 2003b) and MINIPAR (Lin, 1998b) for dependency-based analysis;
- WordNet glosses for transformations involving the additional information.

2.4.2 Sentence Transformations

A similar idea of using a parallel corpus of original and manually adapted texts for learning the transformations which are necessary for an automatic simplification of texts was

used by Petersen and Ostendorf (2007), Gasperin et al. (2009), and Drndarević and Saggion (2012). This time, the focus of the studies was on specific sentence transformations such as splitting and deletion. Although they cannot be directly compared as they were performed on the corpora in different languages and for different target populations, they reveal some interesting phenomena which seem to be independent of the target population and language. Experiments presented in Chapter 4 were mainly inspired by those three previous studies (Petersen and Ostendorf, 2007; Gasperin et al., 2009; Drndarević and Saggion, 2012). Table 2.5 provides a quick overview of the differences and similarities among those three studies.

Table 2.5: Studies on necessary sentence transformations for ATS

	Petersen-07	Gasperin-09	Drndarevic-12
Language	English	Portuguese	Spanish
Target	Language learners	Low literacy	People with ID
Text genre	News	News	News
# of sentences	2588	2685	246
Sent. splitting	Yes	Yes	No
Sent. deletion	Yes	No	Yes
Classifier	C4.5 decision tree	SMO (SVM)	SVM

The columns 'Petersen-07', 'Gasperin-09', and 'Drndarevic-12' represent the studies by Petersen and Ostendorf (2007), Gasperin et al. (2009), and Drndarević and Saggion (2012).

In all three cases, the authors were interested in developing a system for automatic simplification of texts for their specific target population. Although the user groups and languages were different, text genre was the same (news articles) and the observed transformations were similar: some sentences or phrases were deleted, long sentences were split into several shorter ones, long descriptive phrases were shortened, etc. The

authors were not interested in changes to vocabulary in any of the three studies, but rather focused on:

1. Differences in part-of-speech usage and phrase types between original and simplified sentences (Petersen and Ostendorf, 2007);

2. Characteristics of sentences which were chosen to be split (Petersen and Ostendorf, 2007; Gasperin et al., 2009);

3. Characteristics of sentences which were deleted (Petersen and Ostendorf, 2007; Drndarević and Saggion, 2012).

Petersen and Ostendorf (2007) used a corpus of 104 original news articles and their abridged versions developed by Literacyworks, which is freely available on the internet[11]. Gasperin et al. (2009) used corpora from two of the main Brazilian newspapers, *Zero Hora* and *Folha de São Paulo*. The first one (a total of 2,116 original sentences) comprises general news articles, while the second one (a total of 569 original sentences) contains texts from the science section. Drndarević and Saggion (2012) used the corpus of news articles obtained from the Spanish news agency Servimedia[12] and compiled under the Simplext project (Saggion et al., 2011).

Petersen and Ostendorf (2007) reported that out of a total of 2,539 original sentences (100%), 30% were dropped, 19% were split (into two or more abridged sentences), 7% were merged (two original sentences merged into one abridged), and 47% of sentences had '1-1' alignment (one original sentence corresponds to one abridged sentence). The

[11] http://literacynet.org/cnnsf/index_cnnsf.html
[12] http://www.servimedia.es/

proportions of deleted, split, and '1-1' aligned sentences in the Simplext corpus (Štajner et al., 2013) were similar (Table 2.6). The only difference was that in the Simplext corpus, the amount of split and deleted sentences was practically the same. The analysis of the corpora used by Gasperin et al. (2009), however, revealed a significant difference in comparison with the other two as there were almost no deleted sentences. This might be interpreted as an interesting difference in simplification strategies when simplifying texts for different target groups. It seems that simplification of texts for language learners and people with intellectual disabilities requires a fair amount of content reduction (reflected in the number of deleted sentences), while simplification for people with low literacy tries to keep all information which was present in the original text.

Table 2.6: Distribution of sentence transformations

	LiteracyWorks	Wikipedia	PorSimples	Simplext
Language	English	English	Portuguese	Spanish
Genre	News	Wikipedia	News	News
Target	Language learners	Various	Low literacy	People with ID
# of sentences	2,588	90,000	2,685	246
Split	18%	11%	29%	23%
Deleted	29%	31%	0.3%	21%
Merged	6%	7%	0.3%	Unknown
'1-1'	46%	51%	70%	55%

The columns 'LiteracyWorks', 'Wikipedia', 'PorSimples', and 'Simplext' represent the following four studies conducted on the corresponding corpora: (Petersen and Ostendorf, 2007), (Coster and Kauchak, 2011b), (Gasperin et al., 2009), and (Štajner et al., 2013)

Coster and Kauchak (2011b) introduced a new dataset for text simplification by aligning the sentences from English Wikipedia[13] (EW) and Simple English Wikipedia[14]

[13]http://en.wikipedia.org/
[14]http://simple.wikipedia.org

(SEW). Simple English Wikipedia offers a similar content as English Wikipedia presented using simpler vocabulary and grammar in order to facilitate its comprehension to children, English language learners, people with low-literacy levels, and other people with reading difficulties.

Sentences from the EW and SEW were automatically aligned. Two human evaluators estimated the automatic sentence alignment in the EW-SEW dataset as correct in 91% of the cases (on a small portion of 100 sentences), while the other 9% was only partially correct. Out of 137,000 aligned sentence pairs, 27% of sentences were identical and they were excluded from further analysis. 23% of the remaining original sentences could not be aligned with any simplified sentence, and 27% of the remaining simplified sentences could not be aligned with any original sentence. Among the remaining sentence pairs, the '1-1' alignment (one original to one simple sentence) was found in 37% of the cases, the '1-2' (one original to two simple sentences) was found in 8% of the cases, and the '2-1' (two original to one simple sentence) alignments were found in 5% of the cases (Table 2.6). That such a great number of simplified sentences which could not be aligned with any original sentence is the consequence of the fact that the texts in the SEW were not made as direct simplifications of the corresponding original articles, rather they were written independently but following the same topic. For the same reason, the number of *deleted* sentences in Table 2.6 is not directly comparable with the number of deleted sentences in the other TS corpora.

Based on word alignment learned using GIZA++ (Och and Ney, 2003), Coster and Kauchak (2011b) focused their study on word transformations and calculated the percentage of sentences which included:

- Rewordings (a normal word is changed to a different simple word): 65%,
- Deletions (a normal word is deleted): 47%,
- Reorderings (non-monotonic alignment): 34%,
- Merges (multiple normal words are condensed to a single simple word): 31%
- Splits (a normal word is split into multiple simple words): 27%.

2.4.3 Analysis of Split Sentences

The analysis of split and unsplit sentences by Petersen and Ostendorf (2007) revealed that, as expected, split sentences are longer, have a greater number and length of S (simple declarative clauses), SBAR (clauses introduced by a subordinating conjunction), NP (noun phrases), VP (verb phrases), and PP (prepositional phrases) than the unsplit sentences. The classification of original sentences into those to be *split* and those to be left *unsplit* was performed by the C4.5 decision tree learner (with 10-fold cross-validation setup) in order to get results which can easily be interpreted, as the main focus was on the analysis rather than the classification itself. Petersen and Ostendorf (2007) used 20 features in total:

- Sentence length (in words).
- Number of adjectives, adverbs, coordinate conjunctions, prepositions, determiners, nouns, proper nouns, pronouns, and verbs.
- Number and average length of S, SBAR, NP, VP, and PP.

The average cross-validation error rate was reported to be 29%. As was expected, sentence length was the most important feature, leading to the two most used rules: (1) to leave sentences with fewer than 19 words unsplit; and (2) to split sentences with more than 24 words. Additionally, Petersen and Ostendorf (2007) reported that S and SBAR were not commonly used features. This was surprising given that the syntactic simplification modules in rule-based TS systems usually use S and SBAR as the main indicators of whether the given sentence should be split or not.

Besides the features used by Petersen and Ostendorf (2007), Gasperin et al. (2009) used 183 additional features based on lexicalised cue phrases (157 features) and the rhetoric relations (26 features) in their classification experiments for Brazilian Portuguese. They achieved a 0.80 F-measure using the SMO classifier (Weka's implementation of the SVM).

2.4.4 Analysis of Deleted Sentences

Classification between deleted and all other original sentences in the study by Petersen and Ostendorf (2007) was also done using the C4.5 rule generator, but this time with a different set of features, as the authors assumed that the reason for deleting sentences would lie in content-based features rather than in syntactic ones. The chosen set of features was the following:

- Position in the document: sentence number, percentage;
- Paragraph number, first or last sentence in paragraph;
- Does this sentence contain a direct quotation?

- Percentage of words that are stop words (according to the WordNet 1.6 list);

- Percentage of content words which have already occurred one, two, three, four, or more times in the document.

Petersen and Ostendorf (2007) reported classifier performance to be a little better than always choosing the majority class (not deleted). The rule with the highest applicability and lowest error rate was the one choosing to keep sentences which fulfil all of the following conditions: position $\leq$ 12, stop words $\leq$ 70%, content words seen once $\leq$ 40%, no content words seen more than five times. Rules for *deleted* sentences were reported to have lower applicability and higher error rates than rules for *kept* sentences.

For the same task of classification between deleted and kept sentences, this time for Spanish, Drndarević and Saggion (2012) used a different set of features, which included:

- Position of the sentence in the text;

- Number of named entities (NE) and numerical expressions (NumExp);

- Number of content words and punctuation tokens;

- Word frequency distribution;

- Various cohesion features.

Their SVM classifier achieved a 0.79 F-measure in a cross-validation setup, outperforming two baselines: *delete the last sentence*, and *delete the last two sentences*.

2.5 Summary

This chapter presented some of the obstacles which complex texts may pose to human comprehension (Section 2.1) and machine processing (Section 2.2). The former was presented from two different angles: (1) the psycholinguistic perspective (Section 2.1) and the perspective of the existing guidelines for producing easy-to-read texts (Section 2.3); and (2) based on the analyses of parallel corpora containing original texts and their manual simplifications aimed at specific target populations (Section 2.4). The previous studies indicated that building a classifier which would decide on whether the original sentence should be split or left unsplit is a much easier task than building a classifier which would decide on whether the original sentence should be removed or kept in the simplified version of the text (Sections 2.4.3 and 2.4.4). However, none of those proposed decision-making systems has yet been included as a module in an automatic text simplification system, probably due to their still unsatisfying classification accuracies.

CHAPTER 3

AUTOMATIC TEXT SIMPLIFICATION

This chapter presents the most influential and the most complete automatic text simplification (ATS) systems proposed until now. Section 3.1 focuses on the ATS systems which use the rule-based approach. Although being the oldest approach in ATS, this approach is still dominant for languages for which there are no large parallel corpora which would enable data-driven approaches. With the emergence of the Simple English Wikipedia, the focus of the ATS systems for English shifted towards data-driven approaches (Section 3.2). In the last year, a new generation of ATS systems has appeared – the hybrid systems (Section 3.3). The aim of those systems is to combine the best of the two previous generations, the rule-based syntactic simplification and the data-driven lexical simplification. This chapter presents the main characteristics, pros and cons, and best examples of each of the aforementioned approaches. It also provides an overview of different evaluation strategies used in automatic text simplification (Section 3.4).

3.1 Rule-Based ATS Systems

The first ATS systems were rule-based, e.g. (Chandrasekar and Srinivas, 1997; Carroll et al., 1998; Siddharthan, 2002). The implementation of systems proposed by Carroll et al. (1998) and Chandrasekar and Srinivas (1997) encompassed only two stages:

analysis and *transformation*. The first stage (*analysis*) used various sentence analysers (taggers, morphological analysers, parsers, finite-state grammars, etc.) to provide a structural description of the input. The simplification was performed in the second stage (*transformation*), based on the sentence description obtained in the first stage. Different motivations for those two studies – simplification of texts for aphasic readers (Carroll et al., 1998), and simplification of long and complex sentences in order to improve their processing by various natural language processing tools (Chandrasekar and Srinivas, 1997) – led to a use of different analysers and different types of simplifications. While Chandrasekar and Srinivas (1997) were only concerned with syntactic simplification, Carroll et al. (1998) proposed a system which performed both syntactic and lexical simplification (Figure 3.1).

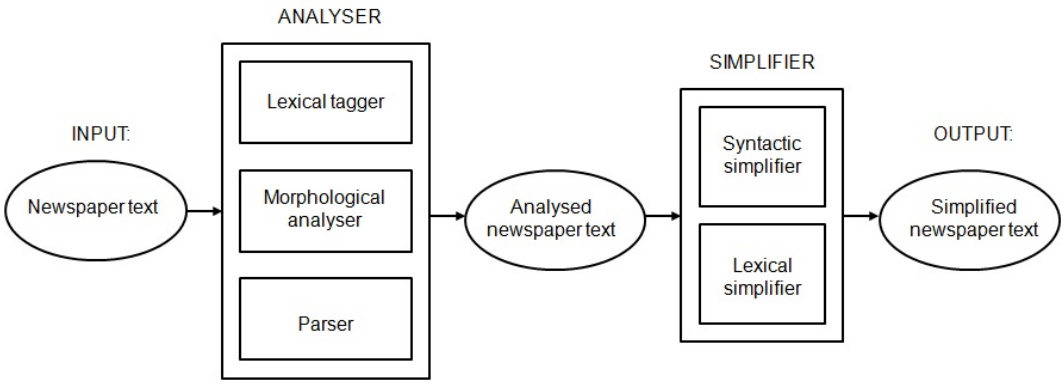


Figure 3.1: System architecture (Carroll et al., 1998)

A few years later, Siddharthan (2002) drew attention to the fact that neither of those two previous systems took into account inter-sentential discourse considerations, which are essential for syntactic simplification if we wish to obtain output which preserves the meaning and coherence of the original text. Therefore, Siddharthan (2002) proposed an

architecture for the ATS system which would, in addition to the *analysis* and *transformation* stages, also contain a third stage – *regeneration* (Figure 3.2).

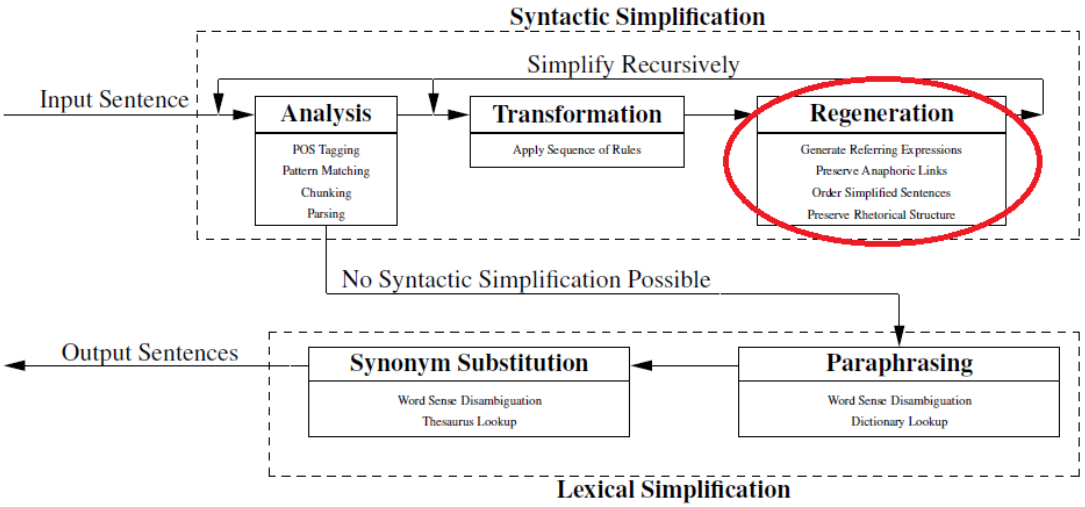


Figure 3.2: Architecture of the ATS system with regeneration stage (Siddharthan, 2002)

A quick overview of the rule-based ATS systems proposed so far, and their main characteristics, is given in Table 3.1. The systems vary according to many criteria, such as the type of the transformation they perform (lexical, syntactic, or both), the language they address, the target population they are intended for, or the number of stages they include (two or three depending on whether they include the *regeneration* stage or not).

Although all ATS systems encompass the analysis stage, they still differ in type of the analysers they use (Table 3.2). Some of them use only chunking and POS tagging (Chandrasekar, 1994; Devlin and Unthank, 2006), some prefer to use only dependencies returned by the parsers (Chandrasekar and Srinivas, 1997; Siddharthan, 2011), while

Table 3.1: Main characteristics of various rule-based ATS systems

Study	Language	Target	Lexical	Syntactic
(Chandrasekar, 1994)	English	MT	No	Yes
(Chandrasekar et al., 1996)	English	NLP tools	No	Yes
(Chandrasekar and Srinivas, 1997)	English	NLP tools	No	Yes
(Carroll et al., 1998)	English	Aphasic	Yes	Yes
(Devlin, 1999)	English	Aphasic	Yes	Yes
(Canning et al., 2000)	English	Aphasic	No	Yes
(Siddharthan, 2002)	English	Various	No	Yes
(Siddharthan, 2006)	English	Various	No	Yes
(Devlin and Unthank, 2006)	English	Aphasic	Yes	No
(Burstein et al., 2007)	English	Learners	Yes	No
(Vickrey and Koller, 2008)	English	SRL	No	Yes
(Bautista et al., 2009)	English	Special	Yes	Yes
(Caseli et al., 2009)	B.Portuguese	Low liter.	Yes	Yes
(De Belder and Moens, 2010)	English	Children	Yes	Yes
(Siddharthan, 2010)	English	Various	No	Yes
(Siddharthan, 2011)	English	Various	No	Yes
(Bott et al., 2012b)	Spanish	People with ID	No	Yes
(Bott et al., 2012a)	Spanish	People with ID	Yes	No
(Aranzabe et al., 2012)	Basque	Various	No	Yes
(Barlacchi and Tonelli, 2013)	Italian	Children	No	Yes
(Brouwers et al., 2014)	French	General	No	Yes

most systems opt for the use of the full parses (Carroll et al., 1998; Vickrey and Koller, 2008; Bautista et al., 2009; Caseli et al., 2009; De Belder and Moens, 2010; Bott et al., 2012b; Barlacchi and Tonelli, 2013; Brouwers et al., 2014).

3.1.1 Lexical Simplification

Lexical simplification traditionally relies on substitution of long and infrequent words with their shorter and more frequent synonyms, following the instructions for easy-to-read texts in the proposed guidelines (Table 2.1, Section 2.3). It encompasses two phases: searching for the synonyms of the 'difficult' word; and choosing the best ('easiest') of those synonyms as a substitute for the 'difficult' word.

Table 3.2: Tools used during the *analysis* stage in different rule-based ATS systems

Study	Tools used during the analysis stage
(Chandrasekar, 1994)	POS tagger + Finite State Grammars (FSG) for chunking
(Chandrasekar and Srinivas, 1997)	Simple dependency representation provided by LTAG (Joshi, 1985; Schabes et al., 1988) + 'supertagging' (Joshi and Srinivas, 1994)
(Carroll et al., 1998)	Lexical tagger (Elworthy, 1994) + morphological analyser (Cunningham et al., 1996) + parser (Briscoe and Carroll, 1993)
(Siddharthan, 2006)	LT Text Tokenization Toolkit (Grover et al., 2000) + WordNet (Miller et al., 1993)
(Devlin and Unthank, 2006)	LT CHUNK POS tagger (Edinburgh University Language Technology Group) + Android Technologies MySQL port of WordNet (Princeton University) + the Irine Phonotactic Dictionary
(Burstein et al., 2007)	WordNet (Miller et al., 1993)
(Vickrey and Koller, 2008)	Parser (Geman and Johnson, 2002)
(Bautista et al., 2009)	WordNet (Miller et al., 1993) + Stanford parser (de Marneffe et al., 2006)
(Caseli et al., 2009)	Lists of simple words (Biderman, 2005; Janczura et al., 2007) + list of discourse markers (Pardo and Nunes, 2006) + parser for Portuguese (Bick, 2000)
(De Belder and Moens, 2010)	WordNet (Miller et al., 1993) + Oxford Psycholinguistic Database (Quinlan, 1992) + Stanford parser (de Marneffe et al., 2006)
(Siddharthan, 2010)	RASP toolkit (Briscoe et al., 2006)
(Siddharthan, 2011)	Typed dependency produced by Stanford parser (de Marneffe et al., 2006)
(Bott et al., 2012b)	Mate-tools parser (Bohnet, 2009)
(Aranzabe et al., 2012)	Morpho-syntactic analyser (Aduriz et al., 1998; Karlsson et al., 1995) + lemmatisation and syntactic function identifier (Aduriz et al., 2003) + multi-word identifier (Ezeiza, 2002) + NER (Alegria et al., 2003)
(Barlacchi and Tonelli, 2013)	TextPro (Pianta et al., 2008) + in-built NER + MaltParser (Lavelli et al., 2009)
(Brouwers et al., 2014)	MELT (Denis and Sagot, 2009) + Bonsai (Candito et al., 2010)

The lexical transformation module in most of the rule-based systems searches for the synonyms of the given 'difficult' word in WordNet (Carroll et al., 1998; Bautista et al., 2009). Burstein et al. (2007) and Bott et al. (2012a) do not depend only on the coverage of a thesaurus, but also look for the synonyms of the 'difficult' word in large corpora using either a statistically-generated word similarity matrix (Lin, 1998a), or a vector space model (Salton et al., 1975). De Belder and Moens (2010) find the possible synonyms in the intersection of those obtained from WordNet and those generated by the Latent Words Language model (LWLM) (Deschacht and Moens, 2009). The LWLM represents a limited form of Word Sense Disambiguation (WSD) which has a goal of eliminating inappropriate substitutions. The following two examples illustrate the role of the LWLM (De Belder and Moens, 2010):

(1) *"Authorities* **employ** *(***use***) various mechanisms to regulate certain behaviors in general."* (De Belder and Moens, 2010)

(2) *"In 2007, about one third of the world's workers were* **employed** *(***used***) in agriculture."* (De Belder and Moens, 2010)

The substitution of *employ* by *use* in the first sentence (1) is performed by both the baseline system (which selects the most frequent synonym given by WordNet) and the system proposed by De Belder and Moens (2010) which additionally performs a form of WSD using the LWLM. The substitution of *employed* by *used* in the second sentence (2) is performed only by the baseline system. The system proposed by De Belder and Moens (2010) does not perform this incorrect substitution due to the added LWLM.

After finding the list of the possible synonyms of the 'difficult' word, the lexical

simplification module selects the best (easiest) one based on: its frequencies (Devlin and Tait, 1998; De Belder and Moens, 2010); its length (Bautista et al., 2009); or the combination of both (Bott et al., 2012a). In English, the frequencies are compared according to the Kučera-Francis frequency (Kučera and Francis, 1967) in a psycholinguistic database (Quinlan, 1992), and in Spanish, based on the Referential Corpus of Contemporary Spanish (*Corpus de Referencia del Español Actual, CREA*)[1].

De Belder and Moens (2010) identified low recall as the main problem of their lexical simplification module; the most difficult words in the texts were often not replaced. Shardlow (2014) classified and quantified the types of errors occurring in the baseline lexical simplification system (which is the basis for all systems presented in this section), drawing attention to the main limitations of the proposed rule-based approaches to lexical simplification. In the examined baseline system, every word with the Kučera-Francis frequency below five was considered as complex (phase 1), WordNet was used for the generation of potential substitutes (phase 2), and the substitution candidates were ranked according to their Kučera-Francis frequency (phase 3). The system did not perform any word sense disambiguation. Shardlow (2014) classified the errors in six categories:

- A complex word misidentified as a simple word (type 2A)
- A simple word misidentified as a complex word (type 2B)
- No substitutions available for the given target word (type 3A)
- No simplifying substitutions available for the target word (type 3B)

[1] http://corpus.rae.es/creanet.html

- The meaning of the sentence changed significantly (type 4)
- A substitute which does not simplify the sentence was chosen (type 5)

The error classification was performed by three human annotators. The most common error type observed was 2B (incorrect classification of simple words as complex words), followed by the 2A type errors (misclassification of complex words as simple words). This led to the highest error rate (65%) in the first phase of the simplification pipeline (complex word identification), followed by 42% error rate in the second phase (generation of possible substitutes), and the lowest error rate (27%) in the last phase (ranking of the substitutes according to the Kučera-Francis frequency).

3.1.2 Syntactic Simplification

The goal of syntactic simplification is to convert structurally complex original sentences into one or more structurally simpler sentences, without changing (or at least, minimally altering) their original meaning. As the focus of this thesis is not on the rule-based syntactic simplification, we will not discuss the implementation details of the previous studies, but rather give an overview of the types of sentence transformation covered and discuss the pros and cons of those approaches. The most frequent types of sentence transformations are presented in Table 3.3.

The coverage of the most frequent sentence transformations (Table 3.3) by various rule-based ATS systems is presented in Table 3.4. The systems cannot be directly compared as they treat different phenomena and languages, do not share the same evaluation dataset nor evaluation strategies. Therefore, we only briefly report on the performance of the most recently proposed ATS systems for English and Spanish.

Table 3.3: Most frequent sentence transformations types

Type	Original	Simplified
Appositions	"John Smith, a New York taxi driver, won the lottery."	"John Smith is a New York taxi driver. John Smith won the lottery."
Relative clauses	"The mayor, who recently got a divorce, is getting married again."	"The mayor recently got a divorce. The mayor is getting married again."
Participial phrases	"The participants (...) will be presented with a book, edited by the town council (...)"	"The participants (...) will be presented with a book. This book is edited by the town council (...)"
Coordinate clauses	"The problem is difficult and there is probably no right answer."	"The problem is difficult. There is probably no right answer."
	"The problem is difficult and has no easy solution."	"The problem is difficult. The problem has no easy solution."
Adverbial clauses	"Needing money to pay my rent, I forced myself to beg my parents."	"I needed money to pay my rent. I forced myself to beg my parents."
Subordinate clauses	"Though all these politicians avow their respect for genuine cases, it's the tritest lip service."	"All these politicians avow their respect for genuine cases. However, it's the tritest lip service."
Passives	"Mary was punched by John."	"John punched Mary."

The examples were taken from the following studies: appositions and relative clauses – (De Belder and Moens, 2010); participial phrases – (Drndarević et al., 2013); coordinate clauses – (Bott et al., 2012b); adverbial clauses, subordinate clauses and passives – (Siddharthan, 2002).

Table 3.4: Sentence transformations covered in rule-based TS systems

System	Language	App.	RelC	PartC	CC	AC	SC	Pass
(Chandrasekar and Srinivas, 1996)	English		Yes			Yes		
(Siddharthan, 2002)	English	Yes	Yes	Yes	Yes	Yes	Yes	Yes
(Vickrey and Koller, 2008)	English	Yes		Yes	Yes	Yes		Yes
(De Belder and Moens, 2010)	English	Yes	Yes		Yes		Yes	
(Bott et al., 2012b)	Spanish		Yes	Yes	Yes	Yes	Yes	
(Aranzabe et al., 2012)	Basque	Yes	Yes			Yes		

App. – Appositions; *RelC* – Relative clauses; *PartC* – Participial clauses; *CC* – Coordinative clauses; *AC* – Adverbial clauses; *SC* – Subordinate clauses; and *Pass* – Passive constructions.

De Belder and Moens (2010) proposed an ATS system for English. Syntactic analysis of the input sentences was done by the Stanford parser, the handcrafted rules were applied recursively in order to generate all possible simplifications of every input sentence, and integer linear programming was used for problem optimisation in order to choose the best simplification (most appropriate for the target population). Three human judges evaluated 100 simplified articles, assessing them as correct or incorrect. The performed simplifications of infix coordination and subordination were rated correct in 70% of the cases, the simplifications of relative clauses were rated correct in 60% and 40% of the cases (depending on the corpus, Wikipedia or Literacyworks), and the simplifications of appositions were rated as correct in 60% and 49% (for Wikipedia and Literacywork, respectively). De Belder and Moens (2010) reported that many incorrect simplifications were made due to parsing errors, where the parser had problems in detecting correct clause boundaries and ends of appositions.

Bott et al. (2012b) proposed an ATS system for Spanish. Syntactic structures were represented by dependency trees produced by the Mate-tools parser (Bohnet, 2009). Based on those structures, the simplification rules were developed within the MATE framework (Bohnet et al., 2000). The system was evaluated on news articles. It reported the following precision (P) and recall (R): P = 39% and R = 66% for relative clauses; P = 64% and R=21% for gerundive constructions; P = 42% and R = 58% for object coordination; and P = 65% and R = 50% for verb phrase and clause coordination.

3.1.3 Regeneration and Text Coherence

Chandrasekar et al. (1996) were the first to raise awareness about maintaining coherence

of simplified texts. They pointed out several issues a proposed TS system should deal with:

- Determining the relative order in simplified sentences;
- Choosing referring expressions (e.g. when splitting a sentence with a relative clause, whether the head noun should be copied into the new sentence – which might lead to an awkward-sounding and repetitive text – or should be replaced with an appropriate pronoun; also deciding on the definite or indefinite article);
- Selecting the right tense for the generated sentences;
- Change or loss of the subtleties of original meaning.

Siddharthan (2002), however, regards the generation of referring expressions as a stylistic problem, not as vital to preserving coherence and meaning of the original texts as, for example, determining the order of simplified sentences, or preserving anaphoric and rhetorical link structure. He proposes methods for addressing these issues and reports very good performance (the referring expressions generator gives correct results in 81% of cases, acceptable results in 13% of cases, and incorrect results in only 7% of cases, while the method for deciding sentence order gave acceptable results in all 100 cases).

3.2 Data-Driven Approaches to ATS

The great expansion of the data-driven approaches to ATS in the last few years is mostly based on the use of English Wikipedia and Simple English Wikipedia. The possibility of

using these two corpora for building an automated text simplification system was investigated by Napoles and Dredze (2010) who showed that statistical classification systems can successfully discriminate texts between these two versions of English Wikipedia.

3.2.1 Lexical Simplification

Yatskar et al. (2010) used edit histories in Simple English Wikipedia to extract lexical simplifications. They proposed two systems, SIMPL and EDIT MODEL, which both significantly outperformed two baselines (RANDOM and FREQUENT) in terms of precision. Both systems were based on unsupervised methods, thus not requiring any human annotation of the data. Furthermore, the two proposed systems seem to be complementary to each other and they are able to extract lexical simplifications not present on the list of simple words and simplifications from Simple Wikipedia (SPLIST) assembled by Spencer Kelly[2] using a combination of dictionaries and manual effort (Yatskar et al., 2010). Some examples of simplifications from the SPLIST and from the systems proposed by Yatskar et al. (2010) are presented in Table 3.5.

Biran et al. (2011) also applied an unsupervised method for learning pairs of complex and simple synonyms from a corpus of texts from the original Wikipedia and Simple English Wikipedia. Unlike Yatskar et al. (2010), Biran et al. (2011) did not use the information from edit histories in Simple English Wikipedia nor did they assume any specific alignment between the articles of the original Wikipedia (EW) and Simple English Wikipedia (SEW). Biran et al. (2011) used SEW only as an in-domain simple cor-

[2]http://simple.wikipedia.org/wiki/User:Spencerk/list_of_straight-up_substitutables;
http://simple.wikipedia.org/wiki/User:Spencerk/multiple_word_translations;
http://simple.wikipedia.org/wiki/User:Spencerk/superbasic_megalist

Table 3.5: Examples of lexical simplifications learned from Simple English Wikipedia

SPLIST	(Yatskar et al., 2010)
at once → immediatelly	stands for → is the same as
as a matter of fact → actually	indigenous → native
to the teeth → heavily	permitted → allowed
was hard up for cash → had no money	concealed → hidden
brush up on → improve	collapsed → fell down
identical to → the same as	annually → every year

pus, in order to extract word frequency estimates (Biran et al., 2011). In that sense, the lexical simplification method proposed by Biran et al. (2011) is more general than the method proposed by Yatskar et al. (2010) as it does not require parallel or comparable corpora of original and simple texts. It only requires two corpora (original and simple) which belong to the same domain. The proposed method consists of two phases: rule-extraction and actual simplification. The rule-extraction phase extracts potential pairs of original and simplified words and the score which indicates the similarity between the words. Additionally, in this stage, the system checks whether the extracted pairs indeed represent the pairs of complex and simpler words, based on two measures: corpus complexity and lexical complexity. In the simplification phase, the system decides which words should be simplified, based on the input sentence and the simplification rules learned in the first stage. The proposed system outperformed the frequency-based baseline (Devlin and Unthank, 2006), in terms of grammaticality, meaning preservation and simplicity.

These were big steps forward for the lexical simplification which was previously based on replacement of difficult words by more common WordNet synonyms or para-

phrases from predefined dictionaries (Section 3.1.1). The new data-driven approaches enabled a better coverage and lexical simplification which is not restricted only to one-to-one word substitution.

3.2.2 Text Simplification as Monolingual Phrase-Based SMT

Specia (2010) approached the problem of ATS in Brazilian Portuguese as a translation problem, translating from original to simplified sentences. She demonstrated that phrase-based SMT works reasonably well even on relatively small parallel corpora (4,483 original sentences and their corresponding simplifications). The corpora consisted of original and manually simplified news texts, aimed at people with basic literacy levels (Caseli et al., 2009; Aluísio and Gasperin, 2010). Specia (2010) also showed that the phrase-table produced during the translation process can adequately cover many types of lexical simplification and simple rewriting.

Another set of recent studies (Coster and Kauchak, 2011a,b; Kauchak, 2013) followed the idea proposed by Specia (2010) and approached the problem of text simplification as an English-to-English translation problem, using the corpus of aligned sentences from the original and simple English Wikipedia. Coster and Kauchak (2011a) extended a statistical phrase-based translation system (Koehn et al., 2007) by adding phrasal deletion to the probabilistic translation model in order to better cover deletion which is a frequent phenomenon in text simplification. Their system used 124,000 aligned sentences for training, 12,000 for development and 1,300 for testing. The proposed system (phrase-based translation system with added phrasal deletion) outperformed the baseline (no simplification performed) and several previously proposed sys-

tems (Cohn, 2009; Knight and Marcu, 2002; Koehn et al., 2007) in terms of the BLEU score (Papineni et al., 2002), and two evaluation measures commonly used in text compression (Clarke and Lapata, 2006): a normalised version of edit distance (SSA), and F1 score calculated over words.

Wubben et al. (2012) performed post-hoc re-ranking on the output sentences (simplification hypotheses) based on their dissimilarity to the input (original sentences) in order to overcome one of the main limitations of the previously proposed systems (Specia, 2010; Coster and Kauchak, 2011a) which are overcautious and leave the original sentence unchanged in most of the cases. Wubben et al. (2012) selected the output that is as different as possible from the original sentence, while at the same time controlling for its adequacy and fluency.

In Chapter 5, we perform several sets of experiments which lead to a better understanding of a PB-SMT approach to text simplification. Based on the experiments in three languages (English, Spanish, and Brazilian Portuguese), we reject the widespread assumption that the success of a PB-SMT approach in ATS largely depends on the size of the training and development datasets, and indicate the more probable causes of the success of such a PB-SMT approach to TS reported in previous studies (Specia, 2010; Coster and Kauchak, 2011a; Kauchak, 2013). We show that the sentence pairs in the training and development datasets can be filtered to improve the 'translation' performance, and we reveal some important differences between cross-lingual MT and the monolingual MT used in TS.

3.2.3 Lexico-Syntactic Simplification

Zhu et al. (2010) proposed a tree-based simplification model, inspired by syntax-based machine translation (Yamada and Knight, 2001). It was the first statistical simplification model which covered splitting, dropping, reordering and substitution. Zhu et al. (2010) paired the corresponding articles in EW and SEW, extracted plain texts, used the Stanford parser (Klein and Manning, 2003b) for sentence boundary detection and tokenisation, applied sentence-level TF*IDF for aligning the corresponding sentence pairs (original and simplified), and trained the tree-based text simplification model (TSM) on the full parse trees. A few examples of the output of their system are presented in Table 3.6.

Table 3.6: Examples of the output of the TS system proposed by Zhu et al. (2010)

Ex.	Version	Sentence
(1)	EW	"**Genetic** engineering has expanded the genes available to breeders to **utilize** in **creating desired** germlines for new crops."
	TSM	"Engineering has expanded the genes available to breeders to **use** in **making** germlines for new crops."
	SEW	"New plants were created with genetic engineering."
(2)	EW	"An umbrella term is a word **that** provides a superset or grouping of related concepts, **also** called a hypernym."
	TSM	"An umbrella term is a word**. A word** provides a superset of related concepts, called a hypernym."
	SEW	"An umbrella term is a word that provides a superset or grouping of related concepts."
(3)	EW	"**Almost as soon as he leaves,** Annius and the guard Publius arrive to **escort** Vitellia to Titus**, who** has now chosen her as his empress."
	TSM	"Annius and the guard Publius arrive to **take** Vitellia to Titus**. Titus** has now chosen her as his empress."
	SEW	"Almost as soon as he leaves, Annius and the guard Publius arrive to take Vitellia to Titus, who has now chosen her as his empress."

*E*W – English Wikipedia (original); *T*SM – tree-based TS model proposed by Zhu et al. (2010); *S*EW – Simple English Wikipedia. All examples are taken from the study by Zhu et al. (2010).

The first example (Table 3.6) illustrates dropping (*Genetic*, and *desired*) and substitution (*utilize* → *use*, and *creating* → *making*) performed by the proposed ATS model. In the second example, the TSM system performs dropping (*also*) and sentence splitting operations. The third example combines sentence splitting with substitution (*escort* → *take*). The TSM system outperformed the standard PB-SMT system in the Moses toolkit trained on the same dataset and several other baselines (Zhu et al., 2010).

Woodsend and Lapata (2011a) followed the idea presented by Yatskar et al. (2010) but instead of just learning lexical simplifications, they used quasi-synchronous grammar (Smith and Eisner, 2006) to learn a wide range of rewriting transformations for text simplification. Woodsend and Lapata (2011a) trained two systems, one using SEW revision histories (REVH), and the other using the simplification corpus made of aligned sentences from EW and SEW (ALIGNED). The proposed systems were fully automated and did not need any human intervention at any moment. The results of the comparison of the output of those systems with the 'gold standard' Simple English Wikipedia articles and two baselines demonstrated that the system creates informative articles, which are simpler to read than the baselines (Woodsend and Lapata, 2011a). As a lexical simplification baseline, the authors used simplification lists made by Spencer Kelly (SPLIST, see Section 3.2.1). The other baseline was the tree-based TS system proposed by Zhu et al. (2010). Two examples of original sentences and their simplified versions produced by various systems are presented in Table 3.7.

Narayan and Gardent (2014) combined a probabilistic module for splitting and deletion with a monolingual translation model for phrase substitution and reordering. The proposed ATS system is based on deep semantic representations (the Discourse Repre-

Table 3.7: Comparison of lexico-syntactic data-driven TS systems

Version	Sentence
EW	"Wonder has recorded several critically acclaimed albums and hit singles, and writes and produces songs for many of his label mates and outside artists as well."
Zhu et *al.*	"Wonder has recorded several praised albums and writes and produces songs. Many of his label mates and outside artists as well."
ALIGNED	"Wonder has recorded several critically acclaimed albums and hit singles. He produces songs for many of his label mates and outside artists as well. He writes."
REVH	"Wonder has recorded many critically acclaimed albums and hit singles. He writes. He makes songs for many of his label mates and outside artists as well."
SEW	"He has recorded 23 albums and many hit singles, and written and produced songs for many of his label mates and other artists as well."
EW	"The London journeys In 1790, Prince Nikolaus died and was succeeded by a thoroughly unmusical prince who dismissed the entire musical establishment and put Haydn on a pension."
Zhu et *al.*	"The London journeys in 1790, prince Nikolaus died and was succeeds by a son became prince. A son became prince told the entire musical start and put he on a pension."
ALIGNED	"The London journeys In 1790, Prince Nikolaus died. He was succeeded by a thoroughly unmusical prince. He dismissed the entire musical establishment. He put Haydn on a pension."
REVH	"The London journeys In 1790, Prince Nikolaus died. He was succeeded by a thoroughly unmusical prince. He dismissed the whole musical establishment. He put Haydn on a pension."
SEW	"The London journeys In 1790, Prince Nikolaus died and his son became prince. Haydn was put on a pension."

EW – English Wikipedia (original); *Zhu et al.* – tree-based ATS system proposed by Zhu et al. (2010); ALIGNED – ATS system by (Woodsend and Lapata, 2011a) trained on aligned sentences; REVH – ATS system by (Woodsend and Lapata, 2011a) trained using SEW revision histories; *SEW* – Simple English Wikipedia. All examples are taken from the study by Woodsend and Lapata (2011a).

sentation Structure – DRS (Kamp, 1981) assigned by Boxer (Curran et al., 2007)). The use of deep semantic representations (instead of sentences and full parse trees used in previous studies) facilitates completion (re-creation of the shared element) in the split sentences and better control over deletion of sentence parts, avoiding deletion of oblig-

atory arguments (Narayan and Gardent, 2014). The following examples of an original sentence (3) and its simplified versions (4) and (5) obtained by the systems proposed by Zhu et al. (2010) and Woodsend and Lapata (2011a), respectively, illustrate the need for the use of deep semantic representation in the splitting operation (Narayan and Gardent, 2014).

(3) "The judge ordered that Chapman should receive psychiatric treatment in prison and sentenced him to twenty years to life."

(4) "The judge ordered that Chapman should get psychiatric treatment. In prison and sentenced him to twenty years to life."

(5) "The judge ordered that Chapman should receive psychiatric treatment in prison. It sentenced him to twenty years to life."

Zhu et al. (2010) fail to copy the shared argument *The judge* to the second sentence (4), while Woodsend and Lapata (2011a) do not replace the antecedent *The judge* with a correct pronoun (5). Those errors are due to the fact that both systems (Zhu et al., 2010; Woodsend and Lapata, 2011a) rely solely on syntax. By contrast, the semantically based system proposed by Narayan and Gardent (2014) correctly copies the shared element *The judge* to the second simplified sentence.

In the next example of an original sentence (6) and its simplified version (7) produced by the system proposed by Zhu et al. (2010), the system incorrectly deletes obligatory argument *gifts* and modifies the sentence meaning to *giving knights and warriors* instead of *giving gifts to knights and warriors* (Narayan and Gardent, 2014).

(6) "Women would also often give knights and warriors gifts that included thyme leaves as it was believed to bring courage to the bearer."

(7) "Women also often give knights and warriors. Gifts included thyme leaves as it was thought to bring courage to the saint."

The probabilistic model trained on semantic representations proposed for handling deletion by Narayan and Gardent (2014) avoids such a deletion of obligatory arguments of a predicate, and thus leads to better meaning preservation.

3.3 Hybrid Approaches to ATS

As already mentioned in Section 3.1.1, the main shortcomings of rule-based approaches to lexical simplification are very limited coverage of the systems (due to the erroneous complex word identification and lack of potential substitutes) and the fact that they are limited to one-to-one word substitution (not able to simplify phrases longer than one word). Those problems were successfully overcome by data-driven approaches presented in Section 3.2.1. The main limitation of data-driven approaches is that they are conditioned with availability of large parallel data (original and simplified texts) and thus only applicable to English at the moment. This, however, should not be taken as a flaw of the proposed methods.

Data-driven approaches to syntactic simplification (Section 3.2.3) still do not seem to significantly outperform rule-based approaches (Section 3.1.2). Although being easier to model (requiring less manual effort and being more adaptive to different genres and languages), data-driven approaches to syntactic simplification seem to produce less

grammatical output, and are still very limited in their scope (e.g. they are not able to model conversion from passive to active voice, a problem successfully solved by rule-based syntactic simplification systems). The only exception to this might be the semantically based ATS system proposed by Narayan and Gardent (2014), which seem to be less erroneous than other (not semantically based) data-driven approaches (Zhu et al., 2010; Woodsend and Lapata, 2011a) and have better coverage than other PB-SMT based systems (Coster and Kauchak, 2011a; Wubben et al., 2012). However, there has been no direct comparison of the data-driven ATS system proposed by Narayan and Gardent (2014) with any rule-based system, which could support this hypothesis.

For the above mentioned reasons, Angrosh and Siddharthan (2014) and Siddharthan and Angrosh (2014) proposed a hybrid TS system which combines a data-driven lexical simplification module with a rule-based syntactic simplification module under the unified framework based on synchronous dependency insertion grammars (Ding and Palmer, 2005). The lexical simplification module is trained on the EW-SEW dataset used previously by Coster and Kauchak (2011a) and Woodsend and Lapata (2011a). The syntactic simplification module covers appositions and relative clauses (26 hand-crafted rules), subordination and coordination (85 rules), voice conversion from passive to active (11 rules), standardisation of quotations into the *"X said: Y"* form (14 rules), and several more transformations. The proposed hybrid system (Angrosh and Siddharthan, 2014; Siddharthan and Angrosh, 2014) outperforms the purely data-driven system (Woodsend and Lapata, 2011a) which covers the same sentence transformations. This is illustrated by examples in Table 3.8.

Table 3.8: Comparison of hybrid and purely data-driven TS systems

Version	Sentence
EW	"Takanobu Komiyama (born October 3, 1984 in Chiba, Japan) is a Japanese football player who currently plays for the J-league team Kawasaki Frontale."
QSG	"His father. Komiyama is a."
Hybrid	"Takanobu Komiyama (born October 3, 1984 in Chiba, Japan) is a Japanese football player. Takanobu Komiyama at present plays for the J-league team Kawasaki Frontale."
SEW	"Takanobu Komiyama (born 3 October 1984) is a Japanese football player. He plays for Kawasaki Frontale."
EW	"Since December 2005 it has also been a candidate for joining the European Union and has applied for NATO membership."
QSG	"Since December 2005, it is a candidate for joining the European Union."
Hybrid	"Since December 2005 it has also been a candidate for joining the European Union. And it has applied for NATO membership."
SEW	"Since December 2005 it has also been a candidate for joining the European Union. It has applied for NATO membership."

EW – English Wikipedia (original); *QSG* – TS system based on quasi-synchronous grammars (Woodsend and Lapata, 2011a) ; *Hybrid* – Hybrid TS system (Siddharthan and Angrosh, 2014); *SEW* – Simple English Wikipedia. Both systems (QSG and Hybrid) are trained on the same dataset (aligned sentences from English Wikipedia and Simple English Wikipedia). All examples are taken from the study by Siddharthan and Angrosh (2014).

3.4 Evaluation of ATS Systems

The ideal way of evaluating an ATS system aimed at providing a more accessible information to a certain target population would be to test its effectiveness on their reading time and comprehension. However, as the access to a specific target population might be difficult, most of the studies perform only the expert (non-final-user) evaluation of their systems, providing the human scores for grammaticality, meaning preservation and simplicity of the system's output. Given that such evaluation is performed only at the sentence level, it is usually combined with the automatic evaluation of simplicity of the

whole texts measured in terms of their readability. Data-driven ATS systems which have the possibility to compare the system's output with the 'gold standard' manual simplification additionally use some of the most common machine translation (MT) evaluation metrics.

3.4.1 Readability Indices for Automatic Evaluation of ATS Systems

Since the 1950s, over 200 readability formulae have been developed for the English language, with over 1,000 studies of their application (DuBay, 2004). Initially, they were used to assess the grade level of textbooks. Later, they were adapted to different domains and purposes, e.g. to measure the readability of technical manuals (Automated Readability Index (Smith and Senter, 1967)), or US healthcare documents intended for the general public (the SMOG grading (McLaughlin, 1969)). The earliest readability formulae were computed only on the basis of average sentence and word length. Due to their simplicity and good correlation with the reading tests, some of them, such as the Flesch-Kincaid Grade Level index (Kincaid et al., 1975) or Flesch Reading Ease score (Flesch, 1949), are still widely in use. For example, the Flesch Reading Ease score "correlates .70 with the 1925 McCall-Crabbs reading test and .64 with the 1950 version of the same test" (DuBay, 2004). Another set of readability formulae are those which depend on average sentence length and the percentage of words which cannot be found on a list of the "easiest" words, e.g. the Dale-Chall readability formulae (Dale and Chall., 1948). Readability formulae, initially intended for assessing English texts, have been adapted to other languages by changing the coefficient before the factors. For example, the Flesch-Douma (Douma, 1960) and Leesindex Brouwer (Brouwer, 1963) formulae

for Dutch represent the adaptations of the Flesch Reading Ease score, while Spaulding's Spanish readability formula (Spaulding, 1956) could be seen as an adaptation of the Dale-Chall formula (Dale and Chall., 1948). The work of van Oosten et al. (2010) showed that readability formulae which are solely based on superficial text characteristics (average sentence and word length) seem to be strongly correlated even across different languages (English, Dutch, and Swedish).

With the recent advances in NLP tools and techniques, new approaches to readability assessment have emerged. Schwarm and Ostendorf (2005), and Petersen and Ostendorf (2009), used statistical language modelling and support vector machines to show that more complex features (e.g. average height of the parse tree, average number of noun and verb phrases, etc.) give better readability prediction than the traditional Flesch-Kincaid readability formula. They based their approach on the texts from Weekly Reader[3], and two smaller corpora: Encyclopedia Britannica and Britannica Elementary (Barzilay and Elhadad, 2003), and CNN news stories and their abridged vesions[4]. Feng et al. (2009) introduced some new cognitively motivated features which should improve automatic readability assessment of texts for people with cognitive disabilities. In addition to three previously used corpora (Weekly Reader, Britannica, and CNN news stories) aimed at second language learners and children, Feng et al. (2009) used a corpus of local news articles which were simplified by human editors in order to make them more accessible for people with mild intellectual disabilities (MID). The texts were further rated for readability by people with MID. The study by Feng et al.

[3]http://www.weeklyreader.com/
[4]http://literacynet.org/cnnsf/

(2009) showed that their newly introduced cognitively motivated features (e.g. entity mentions, lexical chains, etc.) are better correlated with the user comprehension than the Flesch-Kincaid Grade Level index (FKGL).

In spite of those findings, most of the existing ATS systems have still been evaluated by using various readability formulae in combination with human judgements of grammaticality and preservation of meaning. Woodsend and Lapata (2011b) evaluated complexity reduction achieved by the proposed ATS system using the Flesch-Kincaid Grade Level index (Kincaid et al., 1975). In their other work, Woodsend and Lapata (2011a) confirmed the results obtained using the FKGL index by comparing them with the Coleman-Liau readability index (Coleman and Liau, 1975), and the Flesch Reading Ease score (Flesch, 1949). Zhu et al. (2010) applied the Flesch readability score in combination with n-gram language model perplexity.

While all of the aforementioned formulae were made for assessing the readability level of English texts, similar studies have started to appear for other languages as well: Swedish (Roll et al., 2007), German (vor der Brück et al., 2008), Portuguese (Aluísio et al., 2010), French (Francois and Watrin, 2011), Italian (DellOrletta et al., 2011), and Basque (Gonzalez-Dios et al., 2014). However, there have been no similar studies for the Spanish language. We address this issue in Chapter 7, adapting several readability formulae for Spanish in a way that allows them to be computed automatically and explore the possibility of using them in automatic evaluation of TS systems. In the same chapter, we also investigate the adequacy of using several readability formulae in English for automatic evaluation of TS systems.

3.4.2 Automatic Evaluation of ATS Systems with MT Metrics

Recently, many studies which propose data-driven ATS systems include an additional assessment of the systems' output by comparing it with the 'gold standard' manual simplifications, borrowing the MT evaluation metrics such as BLEU (e.g. (Specia, 2010; Zhu et al., 2010; Woodsend and Lapata, 2011a; Coster and Kauchak, 2011a; Wubben et al., 2012; Feblowitz and Kauchak, 2013; Narayan and Gardent, 2014; Vu et al., 2014)), TERp (e.g. (Woodsend and Lapata, 2011a; Vu et al., 2014)), or NIST (e.g. (Specia, 2010; Zhu et al., 2010)).

BLEU (Papineni et al., 2002) is the most widely used MT evaluation metric which measures similarity between the system's output and a human reference. It is based on exact n-gram matching and heavily penalises the reordering of words and the shortening of sentences. NIST (Doddington, 2002) is, like BLEU, based on exact n-gram matching, with the difference that it gives different weights to different n-grams (depending on how likely they are to occur) and that its brevity penalty is less severe (small differences in the length of the system's output and the human reference do not impact the overall score as much as in BLEU). TERp (Snover et al., 2009) measures the number of 'edits' needed to transform the MT output (simplified version of the original sentence in our case) into the reference translation (original sentence in our case). TERp is an extension of TER – Translation Edit Rate (Snover et al., 2006) that utilizes phrasal substitutions (using automatically generated paraphrases), stemming, synonyms, relaxed shifting constraints and other improvements (Snover et al., 2009). The higher the value of TERp (and each of its components), the less similar the original sentence is to its corresponding

simplified sentence.

Zhu et al. (2010) argued that BLEU is not a good measure of systems' performances if the systems perform different simplification operations (e.g. the ATS system modelled by PB-SMT and the tree-based ATS system which performs sentence splitting, dropping, substitution, and reordering), as it is known that "BLEU does poorly at comparing systems with radically different architectures and is most appropriate when evaluating incremental changes with similar architectures." (Jurafsky and Martin, 2008). In Chapter 5, we further explore this question, showing that BLEU is not an adequate measure even for comparing several ATS systems which share similar architecture (PB-SMT), as it mostly reflects the similarity between the original sentences and the 'gold standard' in the test set, and not the success of the actual system.

3.4.3 Human Evaluation of ATS Systems

The output of ATS systems is commonly evaluated by human judgements of its grammaticality (fluency), meaning preservation (adequacy) and simplicity, e.g. (Wubben et al., 2012; Feblowitz and Kauchak, 2013; Coster and Kauchak, 2011a; Angrosh and Siddharthan, 2014). Fluency measures grammatical correctness of the output, simplicity measures how simple the output is, and meaning preservation measures how well the meaning of the simplified sentence corresponds to the meaning of the original sentence. All three scores are usually given on a five-point Likert scale. The exceptions to this are the studies by Narayan and Gardent (2014), with a 0–5 scale, and Biran et al. (2011) who use a 1–3 scale for grammaticality and 0–1 scale for meaning preservation and simplicity. In all cases, the higher score indicates the better output.

In Chapter 6, we propose the Information Relevance (IR) score which should replace the meaning preservation and simplicity scores in the evaluation of ATS systems which perform significant content reduction (dropping parts of the original sentences during simplification). Unlike the meaning preservation score, the information relevance (IR) score does not penalise the elimination of sentence parts; the IR score penalises the elimination of relevant information (which leads to loss of, or change in, original meaning) and rewards the elimination of irrelevant information (which leads to increased simplicity of the sentence).

3.5 Summary

This chapter presented various approaches to automatic text simplification and identified main pros and cons of each of them. The widely spread rule-based approaches (Section 3.1) seem to better address syntactic simplification than the purely data-driven approaches (Section 3.2). In lexical simplification, however, data-driven approaches lead to better coverage and less erroneous output than the rule-based approaches. This led to the recent emergence of the hybrid approaches to ATS which combine data-driven approaches to lexical simplification with rule-based approaches to syntactic simplification (Section 3.3). Evaluation of ATS systems usually combines human assessment of the output (in terms of grammaticality, meaning preservation and simplicity) with automatic evaluation of simplicity (readability indices) or closeness to a 'gold standard' (automatic MT evaluation metrics). The evaluation of ATS systems is still not well-established, and evaluation strategies differ from one system to another thus not allowing a fair comparison between different systems (Section 3.4).

CHAPTER 4

TEXT SIMPLIFICATION DECISIONS

In this chapter, we address the problems of classification of original sentences into: (1) those to be eliminated and those to be kept; and (2) those to be split and those to be left unsplit. As already mentioned in Chapter 2, those issues have already been tackled by several studies. Petersen and Ostendorf (2007) addressed both problems in English, Gasperin et al. (2009) focused only on the second classification problem (2) in Brazilian Portuguese, while Drndarević and Saggion (2012) addressed only the first classification problem (1) in Spanish.

Our focus is on those two classification problems in Spanish. We propose a novel set of features and suggest the use of rule-based and tree-based classifiers instead of the traditionally used SVM classifiers in both classification tasks. We also address some issues which, to the best of our knowledge, have never been raised before (in any language): to which extent the size of the training set and the type and purpose of the simplification applied influence the classification results, and whether the classifiers trained on one type of corpus can be successfully applied to another corpus (aimed at different target populations and consisting of texts from different genres).

4.1 Motivation

Due to the scarcity and limited size of the parallel TS corpora in all languages except English, most of the ATS systems are still rule-based. Such systems usually consist of two main components, a lexical simplification module and a syntactic simplification module. A modular approach to text simplification is also present in hybrid ATS systems (which consist of a data-driven lexical simplification module and a rule-based syntactic simplification module) and even in some purely data-driven approaches which use separate modules for sentence splitting, dropping of sentence parts, reordering, etc. (Chapter 3). A sentence decision module which can classify original sentences into those to be *split* and those to be left *unsplit* could enhance the performance of those systems by filtering out the sentences which do not need to be sent to the sentence splitting module. The first steps towards building such a module for English and Brazilian Portuguese have already been made by Petersen and Ostendorf (2007) and Gasperin et al. (2009). Petersen and Ostendorf (2007) reported an average error rate of 29% for a classifier in English, based on the C4.5 decision tree learner, while Gasperin et al. (2009) achieved an F-measure of 0.80 using the SVM classifier for Brazilian Portuguese. To the best of our knowledge, this classification problem has never been addressed for Spanish before.

The importance of content reduction in text simplification was already emphasised in several studies (Bautista et al., 2011; Saggion et al., 2011). As already mentioned in Section 2.1, certain audiences (e.g. people with intellectual disabilities) have problems in processing large amounts of information. Although deletion of entire sentences is a quite common simplification operation performed by human editors (Petersen and

Ostendorf, 2007; Drndarević and Saggion, 2012), so far there have been no ATS systems which perform this operation automatically.[1] Our evaluation of the ATS system for Spanish built under the Simplext project (Drndarević et al., 2013) indicated the lack of a content reduction module as the main reason for the system's performance being far beyond the human simplification. A sentence decision module which can classify original sentences into those to be *deleted* and those to be *kept*, used as an initial step, could significantly enhance the performance of any text simplification system. The first steps towards building such a module for Spanish were made by Drndarević and Saggion (2012) who obtained an F-measure of 0.79 using an SVM classifier.

One of the main problems for building the above mentioned classifiers is a very limited amount of training data. This is the consequence of the scarcity and the very small sizes of the parallel TS corpora aimed at specific target populations. Therefore, we also explore to which extent the size of the training set and the type of the simplification applied influence the classification results, and whether the classifiers trained on one type of corpus can be successfully applied to another corpus (aimed at different target populations and consisting of texts from different genres). To the best of our knowledge, there have been no similar studies in any language.

4.2 Methodology

The corpora, features, and experimental settings are presented in the next three subsections.

[1] Some data-driven ATS systems (Coster and Kauchak, 2011a; Zhu et al., 2010; Woodsend and Lapata, 2011a) perform a very limited content reduction by deleting some short phrases within a sentence. None of them, however, deletes complete sentences.

4.2.1 Corpora

Both sets of experiments (on sentence splitting and on sentence deletion) were conducted on two sentence-aligned text simplification corpora aimed at two different target populations and containing texts of different genres.

The **Simplext corpus** consists of 200 original news articles in Spanish, provided by the Spanish news agency Servimedia[2], and their simplified versions compiled under the Simplext project (Saggion et al., 2011). Simplification was performed manually by trained human editors, familiar with the particular needs of the target group (people with cognitive disabilities) and following a series of easy-to-read guidelines suggested by Anula (2007). The corresponding pairs of original and simplified texts were first sentence aligned using the alignment tool specially built for this purpose (Bott and Saggion, 2011) and then manually post-edited in order to correct sentence alignment where necessary.

The **FIRST corpus** comprises 25 original texts and their corresponding manually simplified versions (a total of 330 original sentences). The texts belong to different genres: literature, news, health, general culture, and instructions. It was compiled under the FIRST project[3] (Orasan et al., 2013). Texts were manually simplified by five experts who have experience in working with people with autism, keeping in mind the particular needs of this target population. We manually aligned the corresponding pairs of original and simplified texts.

Three main sentence transformations present in both corpora are the following[4]:

[2]http://www.servimedia.es/
[3]http://first-asd.eu/
[4]Both corpora also contained several cases of *merged* sentences ('2-1'). In those cases, two original

1. The original sentence is neither split nor deleted ('1-1' alignment)

2. The original sentence is split into two or more sentences ('1-n' alignment)

3. The original sentence is completely deleted ('1-0' alignment)

The distribution of all three types of sentence transformations across the corpora is presented in Table 4.1. Examples of each sentence transformation are given in Table 4.2.

Table 4.1: Corpus analysis: Sentence transformations

Corpus	'1-0'	'1-n'	'1-1'	Total
Simplext	186 (17%)	358 (32%)	566 (51%)	1100 (100%)
FIRST	41 (12%)	70 (21%)	219 (66%)	330 (100%)

4.2.2 Features

All sentences were parsed with the Connexor's Machinese syntax parser[5], and 23 features (Table 4.3) were automatically extracted using the parser's output. Three sets of features were considered: POS frequencies, syntactic features, and two additional features (*sent* and *word*). The use of the first and second set of features was inspired by the syntactic concept of the projection principle which states that "lexical structure must be represented categorically at every syntactic level" (Chomsky, 1986). This implies that the number of nouns in a sentence is proportional to the number of noun phrases in that sentence, the number of verbs in a sentence is related to the number of clauses and

sentences were merged into one simplified sentence. During that process, many sentence parts were deleted from the original sentences, keeping only the most necessary piece of information in the merged simplified sentence. Such sentences were excluded from our experiments.

[5] www.connexor.com

Table 4.2: Examples of sentence transformations

Type	Original	Simplified
'1-1'	*Abre en Madrid su primera sucursal el mayor banco de China y del Mundo.* (Opens in Madrid its first branch the biggest bank of China and the World.)	*El banco mas importante de China y del mundo abre una oficina en Madrid.* (The most important bank of China and the world opens an office in Madrid.)
'1-n'	*El ICBC ha abierto ya 203 sucursales en un total de 28 países de todo el mundo, también en España desde este lunes.* (The ICBC has opened 203 branches in a total of 28 countries around the world, also in Spain since this Monday.)	*El Banco de China tiene oficinas en muchos países del mundo. Ahora, también tiene una oficina en España.* (The Bank of Chine has offices in many countries around the world. Now it also has an office in Spain.)
'1-n'	*Arranca la liga masculina de Goalball, el único deporte específico para ciegos.* (Starts the men's league of Goalball, the only specific sport for the blind.)	*Comienza la liga masculina de Goalball. El Goalball es el único deporte específico para ciegos.* (Begins the men's league of Goalball. Goalball is the only specific sport for the blind.)
'1-0'	*Como muestra de su envergadura, según datos de 2009, el ICBC tenía en nómina a un total de 386.723 empleados, sólo en China, en un total de 16.232 sucursales.* (As a sign of its size and according to data from 2009, the ICBC had a total of 386,723 employees in China only, in 16,232 branches.)	

verb phrases, etc. (Štajner et al., 2012). Some of the features which belong to the first two groups of features (Table 4.3) have already been used by Petersen and Ostendorf (2007), and Gasperin et al. (2009), addressing the problem of sentence splitting decisions in English and Brazilian Portuguese. The two additional features (*sent* and *word*) were inspired by the work of Drndarević and Saggion (2012) on sentence deletion decisions in Spanish.

Table 4.3: Features

Group	Code	Feature
(I) POS tags	*v*	verb
	ind	indicative
	sub	subjunctive
	imp	imperative
	inf	infinitive
	pcp	participle
	ger	gerund
	adj	adjective
	adv	adverb
	pron	pronoun
	det	determiner
	n	noun
	prep	preposition
	cc	coordinate conjunction
	cs	subordinate conjunction
(II) Syntactic	*main*	head of the verb phrase
	premark	preposed marker
	premod	pre-modifier
	postmod	post-modifier
	nh	head of the noun phrase
	advl	head of the adverbial phrase
(III) Other	*sent*	position of the sentence in the text
	words	number of words in the sentence

4.2.3 Experimental Setup

All classification experiments were performed in Weka[6] (Ian H. Witten, 2005; Hall et al., 2009), using three classification algorithms: the Weka implementation of the C4.5 decision tree learner – J48 (Quinlan, 1993), the JRip rule induction algorithm (Cohen, 1995), and the Weka implementation of Support Vector Machines (SVM) – SMO

[6]http://www.cs.waikato.ac.nz/ml/weka/

(Keerthi et al., 2001; Platt, 1998) with no standardisation or normalisation of the data.[7]

The CfsSubsetEval attribute selection algorithm (Hall and Smith, 1998) implemented in Weka was used to select a subset of best features, after which the classification algorithms were applied to both – the whole feature set (*all*), and to the 'best' subset of features returned by the CfsSubsetEval algorithm (*best*). The CfsSubsetEval attribute selection algorithm uses a correlation-based approach to the feature selection problem, following the idea that "good feature sets contain features that are highly correlated with the class, yet uncorrelated with each other" (Hall, 1999). When compared with a wrapper, the CfsSubsetEval gives similar results to the wrapper and even outperforms the wrapper on small datasets (Hall, 1999).

4.3 Sentence Elimination

The analysis of sentence transformations showed that 17% and 12% of the sentences were eliminated in the Simplext and FIRST corpora in the process of manual simplification (Table 4.1). Therefore, automatic detection of sentences to be deleted would be an important step in automatic text simplification. This problem was previously addressed by Drndarević and Saggion (2012) for Spanish, and by Petersen and Ostendorf (2007) for English.

Drndarević and Saggion (2012) trained the SVM classifier (Li et al., 2002) on the sentence pairs from the first 37 text pairs in the Simplext corpus. They borrowed features from text summarisation and added new ones (e.g. position of the sentence in the

[7]The SMO algorithm with normalisation of the data and the SMO algorithm with standardisation of the data always performed equally well as or worse than the SMO version with no standardisation and normalisation.

text, and number of named entities, numerical expressions, content words and punctuation tokens). Their classification system achieved a 0.79 F-measure, outperforming two baselines: the one that deletes the last sentence, and the other that deletes final two sentences in each document.

Petersen and Ostendorf (2007) trained the C4.5 rule generator on the sentence pairs from the Literacyworks website[8] (aimed at language learners) using content-based features: position of the sentence in the document, paragraph number, whether the sentence is the first or last in the paragraph, does the sentence contain direct quotation, percentage of stop words in the sentence, and percentage of content words which have already occurred in the text. The classifier performance was reported to be "little better than always choosing the majority class (not dropped)" (Petersen and Ostendorf, 2007).

4.3.1 Experiments

We first performed the experiments on 248 sentence pairs from 37 text pairs in the Simplext corpus using 10-fold cross-validation, in order to be comparable with the experiments of Drndarević and Saggion (2012) who used exactly the same dataset and cross-validation setup. The goal was to test the success of the new set of features and classification algorithms.

Next, we conducted the same experiments on five different training sets, all of them being a certain subset of the initial two corpora (Simplext and FIRST). The goal was to investigate the impact of: (1) the size of the datasets, and (2) the type of the simplification performed (different target audiences and different text genres), on the success of

[8] http://literacynet.org/cnnsf/index_cnnsf.html

the classification task. In the first case, the same experiments were performed on four datasets of varying size, all subsets of the Simplext dataset (*Simplext-d1*, *Simplext-d2*, *Simplext-d3*, and *Simplext-d4*). In the second case, the same experiments were performed on two datasets of the same size, one subset of the Simplext dataset (*Simplext-d1*), and the other subset of the FIRST dataset (*FIRST-d*). All training and test datasets had the same ratio of *deleted* vs. *kept* sentences in order to make the experiments as comparable as possible.

Finally, we tested the possibility for adaptation of learnt sentence decisions to different text genres and target populations by training the classifiers on one corpus and testing them on another. We used five different test sets, depending on the specific task and the dataset they were trained on. Those classifiers trained on the subsets of the Simplext corpus (*Simplext-d1*, *Simplext-d2*, *Simplext-d3*, and *Simplext-d4*) were tested on the subsets of the FIRST corpus (*dTest-F* and *FIRST-d*); those classifiers trained on the FIRST corpus (*FIRST-d*) were tested on the subsets of the Simplext corpus (*dTest-S* and *dTest-SL*). In none of the experiments did the test set contain any sentences present in the dataset on which the classifiers were trained. The sizes of all datasets are presented in Table 4.4.

4.3.2 Comparison with the State of the Art

The first set of experiments aimed to investigate the success of the newly proposed features and classification algorithms for the task of classifying original sentences into those to be *deleted* and those to be *kept*. Three classification algorithms (SMO, JRip, and J48) were used on both – the entire set of features (*all*), and only the subset of

Table 4.4: Size of the datasets used in the first set of classification experiments

Type	Name	Corpus	Deleted	Kept	Ratio
Cross-validation	Simplext37	Simplext	51	197	0.26
Training	FIRST-d	FIRST	32	215	0.15
	Simplext-d1	Simplext	32	215	0.15
	Simplext-d2	Simplext	64	430	0.15
	Simplext-d3	Simplext	96	645	0.15
	Simplext-d4	Simplext	128	860	0.15
Test	dTest-F	FIRST	9	60	0.15
	dTest-S	Simplext	9	60	0.15
	dTest-C	Combined	9	60	0.15
	FIRST-d	FIRST	32	215	0.15
	dTest-SL	Simplext	32	215	0.13

The column *Ratio* represents the ratio between *deleted* and *kept* sentences in each dataset. The combined test set (*dTest-C*) contains 5 *deleted* sentences from the Simplext corpus, 4 *deleted* sentences from the FIRST corpus, 30 *kept* sentences from the Simplext corpus, and 30 *kept* sentences from the FIRST corpus. The *dTest-SL* dataset comprises of sentences present in the *Simplext-d4* dataset but not present in any other Simplext training set (*Simplext-d1*, *Simplext-d2*, *Simplext-d3*). The *FIRST-d* dataset is used as a training set in some experiments, and as a test set in others.

best features (*best*) returned by the CfsSubsetEval attribute selection algorithm. All experiments were trained on the dataset previously used by Drndarević and Saggion (2012), enabling direct comparison with the state of the art. The results are presented in Table 4.5.

The performance of the SVM using all features was comparable to the results reported by Drndarević and Saggion (2012). The JRip and J48 algorithms outperformed the SVM classifier reported by Drndarević and Saggion (2012) in both feature set-ups. The greatest improvements were achieved in terms of precision in classifying *deleted* sentences (P = 0.88 for JRip (all), and P = 0.81 for JRip (best)) and recall in classifying *kept* sentences (R = 0.99 for JRip (all), and R = 0.98 for JRip (best) and J48 (best)). This

Table 4.5: Classification into *deleted* and *kept* sentences (10-fold cross-validation)

Method	Deleted			Kept			Overall
	P	R	F	P	R	F	F
SVM*	0.42	0.26	0.30	0.86	0.89	0.87	0.79
DeleteLast*	0.27	0.20	0.23	0.81	0.86	0.84	0.73
Delete2Last*	0.31	0.46	0.37	0.84	0.74	0.79	0.68
SMO (all)	0.42	0.20	0.27	0.82	0.93	0.87	0.74
SMO (best)	0.00	0.00	0.00	0.79	1.00	0.88	0.70
JRip (all)	**0.88**	0.29	0.44	0.84	**0.99**	0.91	**0.81**
JRip (best)	**0.81**	0.25	0.39	0.84	**0.98**	0.90	**0.80**
J48 (all)	0.49	0.47	0.48	0.86	0.87	0.87	0.79
J48 (best)	**0.75**	0.23	0.36	0.83	**0.98**	0.90	**0.79**
KeepAll	0.00	0.00	0.00	0.79	1.00	0.88	0.70

Results of the methods marked with an '*' are taken from the study by Drndarević and Saggion (2012). The best results which outperform the state of the art (Drndarević and Saggion, 2012) are presented in bold.

is particularly important in the context of TS as deletion of the sentences which should be *kept* can deteriorate coherence of the text and lead to a loss of important information. In TS, *deleted* sentences misclassified as *kept* only lead to less content reduction but they cannot deteriorate coherence of the text or lead to a loss of relevant information. Therefore, the sentence decision module (which would delete irrelevant sentences) should ideally achieve a perfect precision (P) on the *deleted* class, and a perfect recall (R) on the *kept* class, in order to be implemented as a part of a TS system.

Petersen and Ostendorf (2007) did not report the actual performance of their classifier; they just stated that it performed "little better than always choosing majority class (not dropped)". Although not directly comparable with the study by Petersen and Ostendorf (2007) because of the different language and corpus, our results for the JRip and J48 classifiers significantly outperformed the majority class (*Keep all*).

Rules returned by JRip algorithm for 'all' features:

```
(sent >= 4) and (postmod <= 3) and (nh <= 4) => deleted
=> kept
```

Rules returned by JRip algorithm for the 'best' subset of features:

```
(sent >= 4) and (noun <= 4) => deleted
=> kept
```

Figure 4.1: *Deleted* vs. *kept* sentences

The CfsSubsetEval attribute selection algorithm returned three features – the sentence position (*sent*), the average number of nouns per sentence (*noun*), and the average number of words per sentence (*words*) – as the *best* subset of the initial 23 features. However, the J48 and JRip classifiers trained using only the *best* features did not significantly outperform the same classifiers trained using the whole set of initial features. The SMO classifier trained using only the *best* features performed significantly more poorly than the SMO classifier trained using the whole set of features. The rules returned by the JRip classifier trained using all features (*all*), and the JRip classifier trained only using the *best* subset of initial features (*best*) are presented in Figure 4.1.

4.3.3 The Impact of Training Size

Given that none of the classification algorithms used on the *best* subset of features has outperformed the same algorithms used on the full set of features, all experiments presented in this section were performed on the full set of features only. Table 4.6 contains the results of 21 experiments varying by size of the training dataset, classification algorithm, and test set. The classifiers trained on the first three training sets (*Simplext-d1*, *Simplext-d2*, and *Simplext-d3*) were tested on two test sets of different sizes, the

smaller *dTest-S* and the larger *dTest-SL*. The classifiers trained on the fourth training set (*Simplext-d4*) were only tested on the smaller test set, as the *Simplext-d4* contains instances of the larger test set (*dTest-SL*).

Table 4.6: The impact of the training size (*deleted* vs. *kept*)

Training set	Algorithm	Size	Test set	Deleted			Kept			Overall
				P	R	F	P	R	F	F
Simplext-d1	SMO	247	dTest-S	0	0	0	0.87	1	0.93	0.81
Simplext-d2	SMO	494	dTest-S	0	0	0	0.87	1	0.93	0.81
Simplext-d3	SMO	741	dTest-S	0	0	0	0.87	1	0.93	0.81
Simplext-d4	SMO	988	dTest-S	0	0	0	0.87	1	0.93	0.81
Simplext-d1	SMO	247	dTest-SL	0	0	0	0.87	1	0.93	0.81
Simplext-d2	SMO	494	dTest-SL	0	0	0	0.87	1	0.93	0.81
Simplext-d3	SMO	988	dTest-SL	0	0	0	0.87	1	0.93	0.81
Simplext-d1	JRip	247	dTest-S	**1**	0.33	0.5	0.91	**1**	0.95	**0.89**
Simplext-d2	JRip	494	dTest-S	0.30	0.33	0.32	0.90	0.88	0.89	0.82
Simplext-d3	JRip	741	dTest-S	0	0	0	0.87	1	0.93	0.81
Simplext-d4	JRip	988	dTest-S	0	0	0	0.87	1	0.93	0.81
Simplext-d1	JRip	247	dTest-SL	0.46	0.19	0.27	0.89	0.97	0.93	0.84
Simplext-d2	JRip	494	dTest-SL	0.33	0.25	0.29	0.89	0.93	0.91	0.83
Simplext-d3	JRip	988	dTest-SL	0.33	0.03	0.06	0.87	0.99	0.93	0.81
Simplext-d1	J48	247	dTest-S	0.50	0.33	0.4	0.90	0.95	0.93	0.86
Simplext-d2	J48	494	dTest-S	0.20	0.44	0.28	0.90	0.73	0.81	0.74
Simplext-d3	J48	741	dTest-S	**1**	0.33	0.50	0.91	**1**	0.95	**0.89**
Simplext-d4	J48	988	dTest-S	**1**	0.11	0.20	0.88	**1**	0.94	**0.84**
Simplext-d1	J48	247	dTest-SL	0.36	0.16	0.22	0.88	0.96	0.92	0.83
Simplext-d2	J48	494	dTest-SL	0.32	0.47	0.38	0.91	0.85	0.88	0.82
Simplext-d3	J48	988	dTest-SL	0.83	0.16	0.26	0.89	0.99	0.94	0.85
Baseline: Keep all				0	0	0	0.87	1	0.93	0.81

All experiments are trained using the whole set of initial features, and tested on two Simplext test sets (*dTest-S* and *dTest-SL*). The smaller dataset (*dTest-S*) consists of 9 *deleted* and 60 *kept* sentences, while the larger dataset (*dTest-SL*) consists of 32 *deleted* and 215 *kept* sentences. The best results which significantly outperformed the baseline, where the precision (P) for the *deleted* sentences is 1, and the recall (R) for the *kept* sentences is 1, are presented in bold.

The baseline which classifies all sentences as *kept* (majority class) is already quite high. It achieves the F-measure of 0.81 due to very unbalanced classes (ratio between *deleted* and *kept* sentences is 0.15 for all training and test sets). However, the JRip

classification algorithm trained on the smallest portion of the data (*Simplext-d1*) and the J48 classification algorithm trained on 741 sentences (*Simplext-d3*) significantly outperform the baseline, achieving the F-measure of 0.89. More importantly, both algorithms achieve a perfect precision on the *deleted* class, and a perfect recall on the *kept* class. Those scores lead to a system which does not classify any *kept* sentence as *deleted*. This is particularly important in the context of TS as already explained in the previous section. These classifiers – which achieve a perfect precision (P) on the *deleted* class, and a perfect recall (R) on the *kept* class – can be implemented in an existing rule-based TS system as a module which would delete irrelevant sentences before sending the rest of the sentences to the simplification modules. Given the small size of the *dTest-S* dataset, the classifiers trained on the first three training sets (*Simplext-d1*, *Simplext-d2*, and *Simplext-d3*) were additionally tested on a larger test set (*dTest-SL*). Regardless of the size of the test sets, the results indicate that more data leads to a worse performance of the JRip classifier. The JRip classifiers trained on the two biggest datasets (*Simplext-d3* and *Simplext-d4*) reach the baseline when tested on the smaller test set. In the case of SMO, the size of the training dataset is irrelevant as all experiments only achieve the baseline. The results of the J48 decision tree algorithm vary depending on the size of the training datasets. The J48 classifiers give the best overall performance (F-measure), precision (P) on the *deleted* class, and recall (R) on the *kept* class. Figure 4.2 presents the rules used by the most successful system (the JRip algorithm trained on 247 sentences from the Simplext corpus).

```
(sent >= 2) and (nh <= 5) => deleted
=> kept
```

Figure 4.2: *Deleted* and *kept* sentences (the best system)

4.3.4 The Impact of the Simplification Purpose and Type

The next goal was to investigate to what extent the simplification purpose and type influences the performance of the first classification task (discriminating between the sentences to be *deleted* and the sentences to be *kept*). The Simplext corpus contains texts adapted to people with cognitive disabilities, while the FIRST corpus contains texts adapted to people with autism spectrum disorders (ASD). The experiments presented in Chapter 7 indicate that the simplification in the Simplext corpus was more severe than in the FIRST corpus, reflecting different needs for text adaptation for those two target populations. Therefore, the goal of this section is to discover how much the success of the first classification task depends on the target population for which the text simplification is performed. In order to explore that question, twelve experiments were conducted, using three different classification algorithms (SMO, JRip, and J48) and two training sets of the same sizes. The first training set is a subset of the Simplext corpus (*Simplext-d1*), and the second training set is a subset of the FIRST corpus (*FIRST-d*). Both training sets have the same ratio of *deleted* and *kept* sentences (0.15). All twelve experiments are performed using the whole set of 23 features.

The built classifiers were first tested on the combined test set (*dTest-C*) which contains an equal number of sentences taken from the FIRST and the Simplext corpora in order to enable a fair comparison of the results. As the results of the classifiers tested on

Table 4.7: The impact of the simplification purpose and type (*deleted* vs. *kept*)

Training set	Algorithm	Test set	Deleted			Kept			Overall
			P	R	F	P	R	F	F
Simplext-d1	SMO	dTest-C	0	0	0	0.87	1	0.93	0.81
FIRST-d	SMO	dTest-C	0	0	0	0.87	1	0.93	0.81
Simplext-d1	SMO	dTest-S	0	0	0	0.87	1	0.93	0.81
FIRST-d	SMO	dTest-F	0	0	0	0.87	1	0.93	0.81
Simplext-d1	JRip	dTest-C	0.19	0.33	0.24	0.89	0.78	0.83	0.75
FIRST-d	JRip	dTest-C	0	0	0	0.87	0.97	0.91	0.79
Simplext-d1	JRip	dTest-S	1	0.33	0.5	0.91	1	0.95	**0.89**
FIRST-d	JRip	dTest-F	0.33	0.11	0.17	0.88	0.97	0.92	0.82
Simplext-d1	J48	dTest-C	0.23	0.33	0.27	0.89	0.83	0.86	0.78
FIRST-d	J48	dTest-C	0	0	0	0.87	0.98	0.92	0.80
Simplext-d1	J48	dTest-S	0.50	0.33	0.4	0.90	0.95	0.93	**0.86**
FIRST-d	J48	dTest-F	0	0	0	0.87	0.97	0.91	0.79
Baseline: Keep all			0	0	0	0.87	1	0.93	0.81

The combined test set (*dTest-C*), which contains: 5 *deleted* sentences from the Simplext corpus, 4 *deleted* sentences from the FIRST corpus, 30 *kept* sentences from the Simplext corpus, and 30 *kept* sentences from the FIRST corpus. Both training sets (*Simplext-d1* and *FIRST-d*) have equal sizes: 32 *deleted* sentences, and 215 *kept* sentences.

the combined test set did not even reach the baseline, the classifiers were additionally tested on their respective test set (those classifiers trained on the *Simplext-d1* training set were tested on the dTest-S dataset, while those classifiers trained on the *FIRST-d* training set were tested on the dTest-F dataset). The results of all twelve experiments are presented in Table 4.7.

The SMO classifier always performs equally as well as the baseline (choosing the majority class), irrespective of the training and test sets. It seems that sentence deletion decisions can be learnt with a greater success on the Simplext corpus than on the FIRST corpus (using the JRip and J48 classifiers). This is particularly pronounced when the classifiers (JRip and J48) are tested within the same corpus they are trained on (using dTest-S and dTest-F test sets for Simplext-d1 and FIRST-d training sets, respectively).

When tested on the combined test set (*dTest-C*), the J48 and JRip classifiers achieve a higher weighted average F-measure if trained on the FIRST corpus than if trained on the Simplext corpus (although still below the baseline). However, those classifiers trained on the Simplext corpus achieve higher precision (P), recall (R), and F-measure (F) on the *deleted* class, which is particularly important in the context of TS.

These results should be regarded with caution given the small sizes of the training sets (only 247 sentences) and particularly test sets (only 69 sentences). In order to allow a fair comparison of the performance of the classifiers trained on two different corpora (Simplext and FIRST), the test set should contain a balanced number of instances from both those corpora. The size of such a test set is limited by the small size of the FIRST corpus, and it cannot be enlarged at this moment (without decreasing the size of the *FIRST-d* training dataset). To the best of our knowledge, there are no other TS corpora in Spanish which could be used to enlarge the test set. Therefore, it is not possible to test whether the same results would hold for larger training datasets.

4.3.5 Adaptation

The compilation of a parallel corpus of original and manually simplified texts for a specific target audience (e.g. people with learning or language disabilities) is both time-consuming and expensive (involving special training for human annotators and adaptation of easy-to-read guidelines to a specific language and target population). Therefore, it would be important to investigate whether the simplification systems (or some of their components) developed for one specific target population and text genre could also be used for text simplification aimed at other target populations and different text types; a

problem never addressed before. This section seeks to fill that gap, exploring whether sentence deletion decisions learned from a parallel corpus compiled for the needs of a specific user group could be used in a TS system aimed at different user groups and text genres.

Thirty classification experiments were performed, using five training sets (*Simplext-d1*, *Simplext-d2*, *Simplext-d3*, *Simplext-d4*, and *FIRST-d*), three classification algorithms (SMO, JRip, and J48), and four test sets (*dTest-F*, *FIRST-d*, *dTest-S*, and *dTest-SL*). Those classifiers trained on the subsets of the Simplext corpus were tested on two datasets from the FIRST corpus (*dTest-F* and *FIRST-d*); the other classifiers trained on the FIRST-d training set were tested on two datasets from the Simplext corpus (*dTest-S* and *dTest-SL*).

The results indicate that the sentence deletion decisions learned on one corpus aimed at a specific target population cannot be successfully applied in a TS system aimed at a different target population (Table 4.8). Out of all 30 experiments, only the JRip classifier trained on the *FIRST-d* dataset and tested on the larger test set (*dTest-SL*) outperformed the baseline. It achieved the F-measure of 0.82, and more importantly, a perfect precision (P) on the *deleted* class and a perfect recall (R) on the *kept* class. In order to minimise the possibility that the good result was due to a lucky random choice of the instances in the test set (*dTest-SL*), the JRip classifier trained on the FIRST corpus (*FIRST-d*) was additionally tested on the *Simplext-d1* dataset, which contains the same number of instances as the *dTest-SL*. The overall F-measure was the same (0.82), although the precision (P) on the *deleted* and the recall (R) on the *kept* class were not perfect (P = 0.5 and R = 0.99, respectively). In that additional experiment, only one (out

Table 4.8: Adaptation of sentence decisions (*deleted* vs. *kept*)

Training set	Algorithm	Test set	Deleted			Kept			Overall
			P	R	F	P	R	F	F
Simplext-d1	SMO	dTest-F	0	0	0	0.87	1	0.93	0.81
Simplext-d2	SMO	dTest-F	0	0	0	0.87	1	0.93	0.81
Simplext-d3	SMO	dTest-F	0	0	0	0.87	1	0.93	0.81
Simplext-d4	SMO	dTest-F	0	0	0	0.87	1	0.93	0.81
Simplext-d1	SMO	FIRST-d	0	0	0	0.87	1	0.93	0.81
Simplext-d2	SMO	FIRST-d	0	0	0	0.87	1	0.93	0.81
Simplext-d3	SMO	FIRST-d	0	0	0	0.87	1	0.93	0.81
Simplext-d4	SMO	FIRST-d	0	0	0	0.87	1	0.93	0.81
FIRST-d	SMO	dTest-S	0	0	0	0.87	1	0.93	0.81
FIRST-d	SMO	dTest-SL	0	0	0	0.87	1	0.93	0.81
Simplext-d1	JRip	dTest-F	0.1	0.33	0.15	0.85	0.55	0.67	0.60
Simplext-d2	JRip	dTest-F	0.06	0.11	0.08	0.85	0.75	0.80	0.70
Simplext-d3	JRip	dTest-F	0.2	0.11	0.14	0.87	0.93	0.90	0.80
Simplext-d4	JRip	dTest-F	0	0	0	0.87	1	0.93	0.81
Simplext-d1	JRip	FIRST-d	0.25	0.50	0.33	0.91	0.77	0.84	0.77
Simplext-d2	JRip	FIRST-d	0.13	0.22	0.16	0.87	0.79	0.83	0.74
Simplext-d3	JRip	FIRST-d	0.18	0.12	0.15	0.88	0.92	0.89	0.80
Simplext-d4	JRip	FIRST-d	0	0	0	0.87	1	0.93	0.81
FIRST-d	JRip	dTest-S	0	0	0	0.87	1	0.93	0.81
FIRST-d	JRip	dTest-SL	**1**	0.03	0.06	0.87	**1**	0.93	**0.82**
Simplext-d1	J48	dTest-F	0.09	0.22	0.13	0.85	0.68	0.76	0.68
Simplext-d2	J48	dTest-F	0.07	0.22	0.10	0.82	0.55	0.66	0.59
Simplext-d3	J48	dTest-F	0.10	0.22	0.14	0.86	0.72	0.78	0.70
Simplext-d4	J48	dTest-F	0.10	0.22	0.14	0.86	0.72	0.78	0.70
Simplext-d1	J48	FIRST-d	0.19	0.28	0.22	0.88	0.82	0.85	0.77
Simplext-d2	J48	FIRST-d	0.21	0.53	0.31	0.91	0.71	0.80	0.73
Simplext-d3	J48	FIRST-d	0.18	0.19	0.18	0.88	0.87	0.87	0.78
Simplext-d4	J48	FIRST-d	0.18	0.22	0.20	0.88	0.86	0.87	0.78
FIRST-d	J48	dTest-S	0	0	0	0.87	1	0.93	0.81
FIRST-d	J48	dTest-SL	0	0	0	0.87	1	0.93	0.81
Baseline: Keep all			0	0	0	0.87	1	0.93	0.81

The *dTest-S* and *dTest-F* test set contain 9 *deleted* and 60 *kept* sentences from the corresponding corpora (Simplext and FIRST, respectively). The *dTest-SL* and *FIRST-d* test sets are larger and contain 32 *deleted* and 215 *kept* sentences from the corresponding corpora (Simplext and FIRST, respectively). All classifiers are trained on the whole set of features.

of 215) *kept* sentences was misclassified as *deleted*.

4.4 Sentence Splitting

After classifying the original sentences into those to be *kept* and those to be *deleted*, the next step is the classification of the *kept* sentences into the ones to be split (*split*), and the ones which do not need to be split (*unsplit*). Similar to the study by Petersen and Ostendorf (2007), *deleted* sentences were excluded from the classification into *split* and *unsplit* sentences, as they might have characteristics of both types of sentences.

4.4.1 Experiments

We first performed the experiments in the cross-validation setup as is common practice (Petersen and Ostendorf, 2007; Gasperin et al., 2009). However, it is important to bear in mind that our results are not directly comparable with those in previous studies as they deal with different languages and different TS corpora (aimed at different target populations). This first set of experiments explored whether the newly proposed set of features and classification algorithms lead to performance comparable to the state of the art.

The subsequent sets of experiments were performed on four different training datasets, all of them being a certain subset of the two initial corpora (Simplext and FIRST). The first goal was to investigate the impact of the size of the training datasets on the success of this classification task. The experiments were performed on three training datasets of varying size, all subsets of the Simplext dataset (*Simplext-s1*, *Simplext-s2*, and *Simplext-s3*) and tested on two test sets of different sizes, both containing only the sentences from the Simplext corpus (not present in the training datasets). The next goal was to explore whether the type of the simplification performed (aimed at different target audiences

and applied to different text genres) impacts the classifiers' performance. The experiments were performed on two training datasets of the same size, one subset of the Simplext dataset (*Simplext-s1*), and the other subset of the FIRST dataset (*FIRST-s*). The performance of the classifiers was tested on a combined test set (*sTest-C*) which contained equal numbers of instances from both corpora (Simplext and FIRST), none of them present in the training datasets. Finally, the possibility of adaptation of sentence splitting decisions to different text genres and target audiences was explored using four training sets (*Simplext-s1*, *Simplext-s2*, *Simplext-s3*, and *FIRST-s*). The classifiers were tested with four different test sets (*sTest-F*, *FIRST-s*, *sTest-S*, *sTest-SL*), depending on the dataset they were trained on. Those classifiers trained on the subsets of the Simplext corpus (*Simplext-s1*, *Simplext-s2*, and *Simplext-s3*) were tested on the subsets of the FIRST corpus (*sTest-F* and *FIRST-s*); those classifiers trained on the FIRST corpus (*FIRST-s*) were tested on the subsets of the Simplext corpus (*sTest-S* and *sTest-SL*). In none of the experiments did the test set contain any sentences present in the dataset on which the classifiers were trained. The sizes of all datasets are presented in Table 4.9.

4.4.2 Comparison with the State of the Art

The results of the 10-fold cross-validation on both corpora are presented in Table 4.10. The CfsSubset attribute selection algorithm returned three features as the *best* subset of features for the Simplext corpus: the sentence position (*sent*), number of gerundive verb forms (*ger*), and number of words (*words*). It returned twelve features as the *best* subset of features for the FIRST corpus: number of verbs (*verb*), indicative verbs (*ind*), subjunctive verbs (*sub*), imperatives (*imp*), gerunds (*ger*), pronouns (*pron*), determiners

Table 4.9: Size of the datasets used in the second set of classification experiments

Type	Name	Corpus	Split	Unsplit	Ratio
Cross-validation	Simplext	Simplext	358	566	0.63
	FIRST	FIRST	70	219	0.32
Training	FIRST-s	FIRST	54	169	0.32
	Simplext-s1	Simplext	54	169	0.32
	Simplext-s2	Simplext	108	338	0.32
	Simplext-s3	Simplext	162	507	0.32
Test	sTest-F	FIRST	16	50	0.32
	sTest-S	Simplext	16	50	0.32
	sTest-C	Combined	16	50	0.32
	FIRST-s	FIRST	54	169	0.32
	sTest-SL	Simplext	54	169	0.32

The column *Ratio* represents the ratio between *split* and *unsplit* sentences in each dataset. Each Simplext training set contains all instances present in any smaller training set (i.e. *Simplext-s2* contains all instances present in *Simplext-s1*; *Simplext-s3* contains all instances present in *Simplext-s2*). The combined test set (*sTest-C*) comprises 8 *split* and 25 *unsplit* sentences from each of the two corpora (Simplext and FIRST). The larger Simplext test set (*sTest-SL*) contains sentences present in *Simplext-s3* but not in *Simplext-s2* and *Simplext-s1*. The *sTest-SL* is never used for testing the classifiers trained on *Simplext-s3*. In some experiments the *FIRST-s* dataset is used as a training set, and in others as a test set.

(*det*), nouns (*noun*), coordinating conjunctions (*cc*), preposed markers (*premark*), heads of the noun phrases (*nh*), and number of words (*words*). The models trained using only the *best* subset of features performed significantly better than the models trained on the whole set of features only in the case of the J48 decision-tree learner trained on the FIRST corpus. In all other cases, the differences in performance were not significant.

Previous works on split decisions (Petersen and Ostendorf, 2007; Gasperin et al., 2009), although not directly comparable to ours because of different languages and corpora, achieved a 29% error rate and an F-score of 0.80, respectively. We therefore consider the performance of our classifiers and set of features acceptable. Table 4.11 contains details of the previous studies (Petersen and Ostendorf, 2007; Gasperin et al.,

Table 4.10: Classification into *split* and *unsplit* sentences (10-fold cross-validation)

Dataset-features	Method	Split			Unsplit			Overall	
		P	R	F	P	R	F	F	ER
FIRST-all	SMO	**0.81**	**0.94**	**0.87**	**0.65**	**0.31**	**0.42**	**0.76**	21%
FIRST-best	SMO	**0.78**	**0.85**	**0.82**	**0.73**	**0.62**	**0.67**	**0.76**	21%
FIRST-all	JRip	0.79	0.85	0.82	0.40	0.31	0.35	0.71	28%
FIRST-best	JRip	0.80	0.82	0.81	0.40	0.37	0.38	0.71	29%
FIRST-all	J48	0.78	0.83	0.80	0.33	0.27	0.30	0.68	31%
FIRST-best	J48	0.81	0.86	0.83	0.45	0.36	0.40	0.73	26%
FIRST-all	Baseline	0.76	1	0.86	0	0	0	0.65	24%
Simplext-all	SMO	0.76	0.84	0.80	0.70	0.59	0.64	0.74	26%
Simplext-best	SMO	0.77	0.86	0.81	0.73	0.59	0.65	0.75	24%
Simplext-all	JRip	**0.84**	**0.84**	**0.84**	**0.75**	**0.76**	**0.75**	**0.81**	19%
Simplext-best	JRip	**0.84**	**0.85**	**0.85**	**0.76**	**0.75**	**0.76**	**0.81**	19%
Simplext-all	J48	0.82	0.83	0.83	0.73	0.71	0.72	0.79	21%
Simplext-best	J48	0.85	0.82	0.83	0.73	0.77	0.75	0.80	20%
Simplext-all	Baseline	0.61	1	0.76	0	0	0	0.46	39%

ER – the average error rate, *P* – precision, *R* – recall, *F* – F-measure. The *Overall F* represents the weighted average F-measure.

Table 4.11: Comparison with the state of the art (*split* vs. *unsplit*)

Study	Split	Unsplit	Total	Ratio	Classifier	F	ER
(Petersen and Ostendorf, 2007)	570	1205	1775	0.47	J48	?	29%
(Gasperin et al., 2009)	728	1328	2056	0.55	SMO	0.80	?
Ours (Simplext)	358	566	924	0.63	J48	**0.80**	**20%**
					SMO	0.75	**24%**
					JRip	**0.81**	**19%**
Ours (FIRST)	79	219	298	0.32	J48	0.73	**26%**
					SMO	0.76	**21%**
					JRip	0.71	**28%**

Ratio – the ratio of split and unsplit sentences; *ER* – the average error rate; *F* – weighted average F-measure. Results of the classifiers which outperformed previous studies are shown in bold.

2009) and our best classifiers. It is worth noting that our datasets contain two (Simplext corpus) and six (FIRST corpus) times fewer instances than the datasets used in the previous studies (Petersen and Ostendorf, 2007; Gasperin et al., 2009). In spite of that, the results obtained are comparable with or better than the state of the art.

The rules returned by the JRip algorithm, trained and tested on the Simplext corpus (in a cross-validation setup), are presented in Figure 4.3. The JRip algorithm trained on the whole feature set returns two rules which solely depend on the number of words in the sentence (*words*) and the position of the sentence in the text (*sent*). The JRip algorithm trained on the *best* subset of features returns four rules which additionally take into consideration the number of nouns (*noun*) and coordinating conjunctions (*cc*) in the given sentence.

Rules returned by JRip algorithm using *all* features:

```
(words >= 29) and (sent <= 4) => split
 => unsplit
```

Rules returned by JRip algorithm using only the *best* features:

```
(words >= 34) and (sent <= 4) => split
(words >= 29) and (sent <= 4) and (words <= 31) => split
(words >= 22) and (sent <= 5) and (noun <= 6) and (cc >= 1) => split
 => unsplit
```

Figure 4.3: *Split* vs. *same* sentences (*Simplext* dataset)

4.4.3 The Impact of Training Size

We investigated the impact of the training size through 48 experiments varying by the size of the training dataset, classification algorithm, test set, and the set of features used.

The first 24 experiments were conducted using the whole set of 23 features (Table 4.12); the next 24 experiments were conducted using only the *best* subset of initial features (*sent*, *ger*, and *words*). In all training and test sets, the ratio of split and unsplit sentences was the same (0.32). All experiments were tested using two Simplext test sets: the smaller *sTest-S* dataset (consisting of 16 *split* and 50 *unsplit* sentences) and the larger *sTest-SL* dataset (consisting of 54 *split* and 169 *unsplit* sentences).

Table 4.12: The impact of training size – all features (*split* vs. *unsplit*)

Training set	Algorithm	Size	Test set	Unsplit			Split			Overall
				P	R	F	P	R	F	F
Simplext-s1	SMO	223	sTest-S	0.77	1	0.87	1	0.06	0.12	0.69
Simplext-s2	SMO	446	sTest-S	0.77	0.94	0.85	0.40	0.12	0.19	0.69
Simplext-s3	SMO	669	sTest-S	0.80	0.88	0.84	0.45	0.31	0.37	0.72
Simplext-s1	SMO	223	sTest-SL	0.78	1	0.88	1	0.11	0.20	0.71
Simplext-s2	SMO	446	sTest-SL	0.78	1	0.88	1	0.11	0.20	0.71
Simplext-s1	JRip	223	sTest-S	0.78	1	0.88	1	0.12	0.22	0.72
Simplext-s2	JRip	446	sTest-S	0.81	0.84	0.82	0.43	0.37	0.40	0.72
Simplext-s3	JRip	669	sTest-S	0.86	0.74	0.80	0.43	0.62	0.51	0.73
Simplext-s1	JRip	223	sTest-SL	0.77	1	0.87	1	0.07	0.14	0.69
Simplext-s2	JRip	446	sTest-SL	**0.89**	0.99	0.94	0.97	**0.61**	0.75	**0.89**
Simplext-s1	J48	223	sTest-S	0.83	0.88	0.85	0.54	0.44	0.48	0.76
Simplext-s2	J48	446	sTest-S	0.80	0.86	0.83	0.42	0.31	0.36	0.71
Simplext-s3	J48	669	sTest-S	0.80	0.72	0.76	0.33	0.44	0.38	0.67
Simplext-s1	J48	223	sTest-SL	0.81	0.77	0.79	0.38	0.44	0.41	0.70
Simplext-s2	J48	446	sTest-SL	0.86	0.99	0.92	0.93	0.52	0.67	0.86
Baseline: Unsplit all				0.76	1	0.86	0	0	0	0.65

All experiments are trained using the whole set of initial features, and tested on two Simplext test sets (*sTest-S* and *sTest-SL*). The smaller dataset (*sTest-S*) consists of 16 *split* and 50 *unsplit* sentences, while the larger dataset (*sTest-SL*) consists of 54 *split* and 169 *unsplit* sentences. The best results are presented in bold.

Irrespective of the feature set used, the performance of the classification algorithms was better on larger datasets (Tables 4.12 and 4.13). The only exceptions to this were the J48 classifier trained on the whole feature set (Table 4.12), and the JRip classifier trained

Table 4.13: The impact of training size – *best* features only (*split* vs. *unsplit*)

Training set	Algorithm	Size	Test set	Unsplit			Split			Overall
				P	R	F	P	R	F	F
Simplext-s1	SMO	223	sTest-S	0.76	1	0.86	0	0	0	0.65
Simplext-s2	SMO	446	sTest-S	0.76	1	0.86	0	0	0	0.65
Simplext-s3	SMO	669	sTest-S	0.81	0.94	0.87	0.62	0.31	0.42	0.76
Simplext-s1	SMO	223	sTest-SL	0.76	1	0.86	0	0	0	0.65
Simplext-s2	SMO	446	sTest-SL	0.76	1	0.86	0	0	0	0.65
Simplext-s1	JRip	223	sTest-S	0.80	0.88	0.84	0.45	0.31	0.37	0.72
Simplext-s2	JRip	446	sTest-S	0.81	0.92	0.86	0.56	0.31	0.40	0.75
Simplext-s3	JRip	669	sTest-S	0.86	0.72	0.78	0.42	0.62	0.50	0.71
Simplext-s1	JRip	223	sTest-SL	**0.87**	**1**	**0.93**	**1**	**0.56**	**0.71**	**0.88**
Simplext-s2	JRip	446	sTest-SL	**0.87**	**1**	**0.93**	**1**	**0.54**	**0.70**	**0.87**
Simplext-s1	J48	223	sTest-S	0.76	1	0.86	0	0	0	0.65
Simplext-s2	J48	446	sTest-S	0.77	0.88	0.82	0.33	0.19	0.24	0.68
Simplext-s3	J48	669	sTest-S	0.81	0.88	0.85	0.50	0.37	0.43	0.74
Simplext-s1	J48	223	sTest-SL	0.76	1	0.86	0	0	0	0.65
Simplext-s2	J48	446	sTest-SL	0.80	0.98	0.88	0.76	0.24	0.37	0.76
Baseline: Unsplit all				0.76	1	0.86	0	0	0	0.65

All classifiers are trained only on the *best* subset of initial features ({*sent*, *ger*, and *words*}), and tested on two Simplext test sets (*sTest-S* and *sTest-SL*). The smaller dataset (*sTest-S*) consists of 16 *split* and 50 *unsplit* sentences, while the larger dataset (*sTest-SL*) consists of 54 *split* and 169 *unsplit* sentences. The best results are presented in bold.

on the *best* feature set (Table 4.13), both tested on the smaller test set (*sTest-S*). However, one can argue that even in those two cases, classifiers trained on the largest training dataset (*Simplext-s3*) had the best performance, as they achieved the highest recall of *split* sentences. In the practical application of sentence splitting decisions in TS, recall of *split* sentences is arguably the most important measure of the system's performance. While misclassification of *unsplit* sentences into *split* sentences can only lead to an oversimplified output, misclassification of *split* sentences into *unsplit* sentences leads to inefficiency on the part of the TS system. The best results (F-measure of 0.89) were achieved by the JRip classifier trained on the *Simplext-s2* dataset using the whole set of

initial features and tested on the larger test set (*sTest-SL*).

4.4.4 The Impact of the Simplification Purpose and Type

Our next goal was to investigate to which extent the simplification purpose and type influences the performance of the second classification task (discriminating between the sentences to be *split* and the sentences to be left *unsplit*). In order to explore that question, we conducted six further experiments, using three different classification algorithms (SMO, JRip, and J48) and two training sets of the same sizes. The first training set was a subset of the Simplext corpus (*Simplext-s1*), and the second training set was a subset of the FIRST corpus (*FIRST-s*). Both training sets had the same ratio of *split* and *unsplit* sentences (0.32). All six experiments were performed using the whole set of 23 features. The built classifiers were tested on the combined test set (*sTest-C*) which contains an equal number of sentences taken from the FIRST and the Simplext corpora in order to enable fair comparison of the results.

The results of this set of experiments are presented in Table 4.14. It seems that sentence splitting decisions can be learnt with greater success from the Simplext corpus than the FIRST corpus. However, these results should be taken with caution given the small sizes of the training sets and particularly test sets. In order to allow a fair comparison of the performance of the classifiers trained on two different corpora (Simplext and FIRST), the test set should contain the same number of instances from both corpora. The size of such a test set is limited by the small size of the FIRST corpus, and it cannot be enlarged at this moment (without decreasing the size of the *FIRST-s* training dataset). To the best of our knowledge, there are no other TS corpora in Spanish which could be

Table 4.14: The impact of the simplification purpose and type (*split* vs. *unsplit*)

Training set	Algorithm	Unsplit			Split			Overall
		P	R	F	P	R	F	F
Simplext-s1	SMO	0.77	1	0.87	1	0.06	0.12	0.69
FIRST-s	SMO	0.78	0.78	0.78	0.31	0.31	0.31	0.67
Simplext-s1	JRip	0.79	0.98	0.87	0.75	0.19	0.30	0.74
FIRST-s	JRip	0.76	1	0.86	0	0	0	0.65
Simplext-s1	J48	0.83	0.86	0.84	0.50	0.44	0.47	0.75
FIRST-s	J48	0.81	0.78	0.80	0.39	0.44	0.41	0.70
Baseline: Unsplit all		0.76	1	0.86	0	0	0	0.65

All experiments were performed using the whole set of features and tested on the combined test set (*sTest-C*), which comprises 16 *split* and 50 *unsplit* sentences (8 *split* and 25 *unsplit* sentences from each of the two corpora, Simplext and FIRST). Both training sets (*Simplext-s1* and *FIRST-s*) have equal sizes: 54 *split* sentences, and 169 *unsplit* sentences.

used to enlarge the test set.

4.4.5 Adaptation

As already mentioned in Section 4.3.5, it is important to investigate whether TS systems (or some of their components) developed for one specific target population and text genre could also be used for text simplification aimed at other target populations and different text types; a problem never addressed before. This section seeks to fill that gap, exploring whether sentence splitting decisions learned from a parallel corpus compiled for the needs of a specific user group could be used for different user groups and text genres.

In order to explore this question, we performed 24 classification experiments using four training sets (*Simplext-s1*, *Simplext-s2*, *Simplext-s3*, and *FIRST-s*) and three classification algorithms (SMO, JRip, and J48). Those classifiers trained on the subsets of

the Simplext corpus were tested on two datasets taken from the FIRST corpus (*sTest-F* and *FIRST-s*); the other classifiers trained on the *FIRST-s* training set were tested on two datasets from the Simplext corpus (*sTest-S* and *sTest-SL*).

Table 4.15: Adaptation of sentence decisions (*split* vs. *unsplit*)

Training set	Algorithm	Test set	Unsplit			Split			Overall
			P	R	F	P	R	F	F
Simplext-s1	SMO	sTest-F	0.76	0.96	0.85	0.33	0.06	0.10	0.67
Simplext-s2	SMO	sTest-F	0.76	0.96	0.85	0.33	0.06	0.10	0.67
Simplext-s3	SMO	sTest-F	0.80	0.96	0.87	0.67	0.25	0.36	0.75
Simplext-s1	SMO	FIRST-s	0.76	0.99	0.86	0	0	0	0.65
Simplext-s2	SMO	FIRST-s	0.77	1	0.87	1	0.06	0.10	0.68
Simplext-s3	SMO	FIRST-s	0.76	0.95	0.85	0.33	0.07	0.12	0.67
FIRST-s	SMO	sTest-S	0.77	0.74	0.75	0.28	0.31	0.29	0.64
FIRST-s	SMO	sTest-SL	**0.83**	0.98	0.90	0.83	**0.37**	0.51	**0.80**
Simplext-s1	JRip	sTest-F	0.77	0.98	0.86	0.50	0.06	0.11	0.68
Simplext-s2	JRip	sTest-F	0.80	0.90	0.85	0.50	0.31	0.38	0.74
Simplext-s3	JRip	sTest-F	0.80	0.88	0.84	0.45	0.31	0.37	0.72
Simplext-s1	JRip	FIRST-s	0.76	0.99	0.86	0	0	0	0.65
Simplext-s2	JRip	FIRST-s	0.76	0.95	0.84	0.20	0.04	0.06	0.65
Simplext-s3	JRip	FIRST-s	0.77	0.89	0.82	0.31	0.15	0.20	0.67
FIRST-s	JRip	sTest-S	0.73	0.86	0.79	0	0	0	0.60
FIRST-s	JRip	sTest-SL	0.76	0.99	0.86	0	0	0	0.65
Simplext-s1	J48	sTest-F	0.78	0.92	0.84	0.43	0.19	0.26	0.70
Simplext-s2	J48	sTest-F	0.77	0.94	0.85	0.40	0.12	0.19	0.69
Simplext-s3	J48	sTest-F	0.80	0.86	0.83	0.42	0.31	0.36	0.71
Simplext-s1	J48	FIRST-s	0.75	0.89	0.81	0.14	0.06	0.08	0.64
Simplext-s2	J48	FIRST-s	0.76	0.95	0.84	0.31	0.07	0.12	0.67
Simplext-s3	J48	FIRST-s	0.76	0.85	0.80	0.28	0.18	0.22	0.66
FIRST-s	J48	sTest-S	0.79	0.76	0.78	0.33	0.37	0.35	0.67
FIRST-s	J48	sTest-SL	0.82	0.93	0.87	0.61	0.35	0.45	0.77
Baseline: Unsplit all			0.76	1	0.86	0	0	0	0.65

The *sTest-S* and *sTest-F* test set contain 16 *split* and 54 *unsplit* sentences from the corresponding corpora (Simplext and FIRST, respectively). The *sTest-SL* and *FIRST-s* test sets are larger and contain 54 *split* and 169 *unsplit* sentences from the corresponding corpora (Simplext and FIRST, respectively). All classifiers are trained on the whole set of features.

The results presented in Table 4.15 indicate that the sentence splitting decisions are

universal, i.e. that they can be successfully learnt from one TS corpus (aimed at a specific target population and dealing with a specific text genre) and applied to other text genres and for other target populations. The applicability of the sentence splitting decisions seem to improve with the larger sizes of the training sets, especially when tested on larger test sets (*sTest-SL* and *FIRST-s*). The SMO classifier trained on the FIRST corpus (*FIRST-s*) achieves the F-measure of 0.80 when tested on the larger test set from the Simplext corpus (*sTest-SL*).

4.5 Summary

This chapter presented a series of experiments which address the problems of sentence deletion and sentence splitting decisions in text simplification. The results indicated the following:

1. The newly proposed feature set (consisting of 23 features that can easily be automatically extracted from the parser's output) lead to performance of the classifiers which is comparable to that of the state of the art in both classification tasks.

2. The JRip classifier (together with the CfsSubsetEval attribute selection algorithm) identified the sentence position (*sent*), number of noun phrases (*nh*) and post-modifiers (*postmod*) as the most relevant features in sentence deletion decision making, and the sentence position (*sent*), and number of words (*words*), nouns (*noun*), and coordinating conjunctions (*cc*) as the main features in sentence splitting decision making.

3. The JRip and J48 classifiers achieve performance equal to or better than the SVM

classifier on both classification tasks.

4. The size of the training set significantly influences classification performance in most cases (the larger the training set, the better the classifier's performance).

5. The performance of the classifiers in both tasks depends on the type of the simplification present in the training set.

6. The sentence deletion decisions trained on one type of TS corpus cannot be successfully applied to different text genres and TS aimed at a different target population.

7. The sentence splitting decisions trained on one type of TS corpus can be successfully applied to different text genres and TS aimed at a different target population.

The results presented in this chapter should be treated with caution, given the very limited sizes of the corpora (especially the FIRST corpus). The scarcity and very limited sizes of the parallel corpora of original texts and their manual simplifications performed by trained human editors are some of the biggest challenges in data-driven text simplification. To the best of our knowledge, the Simplext and FIRST corpora are the only existing parallel TS corpora in Spanish. Therefore, at this moment, it is not possible to investigate whether the same findings would hold for larger corpora.

CHAPTER 5

PHRASE-BASED SMT MODELS FOR TEXT SIMPLIFICATION

In the last few years, a growing number of studies have addressed the text simplification (TS) task as a monolingual machine translation (MT) problem of translating sentences from 'original' to 'simple' language. Several studies reported promising results using standard phrase-based statistical machine translation (PB-SMT) for this task, but they did not try to seek the reasons behind the success of their systems. The goal of this chapter is to investigate several important issues in MT-based text simplification: (1) the impact of the size of the training and development datasets on the system's performance; (2) the impact of the type of the simplification on the system's performance; (3) the impact of the type of the datasets (parallel or comparable) on the system's performance; and (4) suitability of the BLEU score for the automatic evaluation of system's performance. To the best of our knowledge, there have been no studies which address those important questions.

5.1 Motivation

In the last few years, there have been several attempts to approach text simplification as a monolingual statistical machine translation (SMT) problem. Instead of translating sentences from one language to another, the goal of text simplification is to translate sen-

tences from 'original' to 'simplified' language. Specia (2010) used phrase-based SMT provided by the Moses toolkit (Koehn et al., 2007) to translate from 'original' to 'simple' sentences in Brazilian Portuguese. In terms of the automatic BLEU evaluation (Papineni et al., 2002), the results were reasonably good (BLEU = 60.75) despite the small size of the corpora (4,483 original sentences and their corresponding simplifications). Coster and Kauchak (2011a) extended a statistical phrase-based translation system by adding phrasal deletion to the probabilistic translation model in order to better cover deletion, which is a frequent phenomenon in text simplification. Their system, trained on 124,000 aligned sentences from English Wikipedia and Simple English Wikipedia, achieved a BLEU score of 60.46 (59.87 on the standard model without phrasal deletion). These studies (Specia, 2010; Coster and Kauchak, 2011a) indicated that the size of the datasets might not be the most important factor for the success of the standard PB-SMT models in text simplification.

5.2 Methodology

The methodology employed was as follows:

- Running the standard PB-SMT experiments on three different datasets and languages.
- Comparing the distributions of sentence-level BLEU scores across three training sets and one additional TS corpus.
- Investigating whether the method of collecting the datasets (from parallel or comparable corpora) influences the results of the standard PB-SMT in text simplifica-

tion.

- Testing how the quality of the data (i.e. the sentence similarity between the original and simplified sentences in the training and development datasets) influences the results of the standard PB-SMT in text simplification.

- Investigating whether the sizes of the training and development datasets influence the results of the standard PB-SMT in text simplification.

The datasets and the experimental setup for the translation experiments are described in the next two sub-sections.

5.2.1 Investigated Datasets and Languages

We used four text simplification corpora in three different languages:

1. **Simplext** – The corpus of original news texts in Spanish and their manual simplifications aimed at people with Down's syndrome (Drndarević et al., 2013; Štajner and Saggion, 2013), containing 925 sentence pairs. Simplification was performed by trained human editors under the Simplext project (Saggion et al., 2011).

2. **PorSimples** – The corpus of original news texts in Brazilian Portuguese and their manual simplifications for people with low literacy levels (Caseli et al., 2009). It contains 4,483 original sentences with two manually simplified versions of each of them, using 'natural' and 'strong' simplifications (depending on the literacy level of the readers). The original sentences and their corresponding 'natural' simplifications of this corpus were used by Specia (2010).

3. **Wikipedia** – The corpus of 137,000 automatically aligned sentence pairs from the comparable articles in English Wikipedia and Simple English Wikipedia, used by Coster and Kauchak (2011a).

4. **EncBrit** – The parallel corpus of original sentences from Encyclopedia Britannica and their manually simplified versions for children (Barzilay and Elhadad, 2003). Given its small size (601 sentence pairs) this dataset was used in the translation experiments only as a test set. In the experiment on assessing the differences in distribution of S-BLEU scores across the training datasets, this corpus was used as an additional dataset which contains strong simplifications (similar to the Simplext data but in another language and for a different target audience).

Here it is important to emphasise that while *Simplext* and *PorSimples* are parallel corpora of original and manually simplified sentences, the *Wikipedia* and *EncBrit* corpora are only comparable. The simplified versions of the articles in *Wikipedia* and *EncBrit* corpora were written independently of the original articles. The sentences in two versions were automatically aligned using the procedures for sentence alignment for monolingual comparable corpora (Barzilay and Elhadad, 2003; Coster and Kauchak, 2011b).

5.2.2 Experimental Setup for the Translation Experiments

In all translation experiments, we used the standard PB-SMT system in the Moses toolkit (Koehn et al., 2007) with the GIZA++ implementation of IBM word alignment model 4 (Och and Ney, 2003), and the refinement and phrase-extraction heuristics described further by Koehn et al. (2003). The systems were tuned using minimum error

rate training (MERT) (Och, 2003). In all experiments, the language model was built using the 3-gram language model with Kneser-Ney smoothing trained with SRILM (Stolcke, 2002) on a 500,000 sentence corpus. The stack size was limited to 500 hypotheses during decoding.[1]

5.2.3 Evaluation

Our first set of translation experiments (Section 5.3) was evaluated automatically using the BLEU score (Section 5.3.1) and manually for the error analysis of the experiments in English and Spanish (Section 5.3.2). The second set of translation experiments (Section 5.5) was evaluated automatically using the BLEU score (Section 5.5.2) and by three human annotators who assessed grammaticality, meaning preservation and simplicity of the generated output (Section 5.6).

In spite of the many objections to using BLEU as an automatic measure of MT systems' performance, BLEU still seems to be more appropriate than other existing measures for the automatic evaluation of MT systems, in this specific task. The reason for this lies in differences between cross-lingual MT and the monolingual MT used for TS. When translating from one language to another, there are many different translations that can be equally good. In that case, it is necessary to have several reference translations and/or some automatic evaluation metric which does not penalise outputs which are worded differently, but still convey the same meaning. When translating from 'original' to 'simplified' language, there is usually only one reference 'translation' (simplified version). A different choice of words or reordering of the clauses, even if it

[1] Our experimental setup is the same as in the previous studies which used PB-SMT for TS (Specia, 2010; Coster and Kauchak, 2011a).

conveys exactly the same meaning, might not be equally suitable for the specific target population. It could even lead to the output being more difficult to understand than the original sentence. Therefore, in the case of MT used for TS, it is necessary to have an evaluation metric which would penalise the output sentences that are different to the given reference even if they are completely grammatical and have the same meaning as the reference 'translation'. Therefore, BLEU seems more suitable than any other MT evaluation metric (e.g. METEOR, TERp) for this task.

5.3 Translation Experiments across the three Corpora

We conducted three MT experiments using the standard PB-SMT system described in Section 5.2.2. The English experiment used the Wikipedia TS corpus for the translation model (TM) and the English part of the Europarl corpora[2] to build the language model (LM). The Spanish experiment used the Simplext corpus to build the TM and the Spanish Europarl for the LM. The Brazilian Portuguese experiment used the PorSimples corpus for the TM and the Lácio-Web corpus[3] in Brazilian Portuguese for the LM. We are aware of the fact that using 'original' (unsimplified) sentences to build the language model probably influences the simplicity of the generated sentences negatively. Ideally, the LM should be trained on 'simple' sentences. However, a large enough corpus of 'simple' sentences exists only in English (the Simple English Wikipedia), while there are no similar corpora in Spanish or Brazilian Portuguese. Therefore, we opted for building the language models on the corpora consisting of 'original' sentences for all

[2]http://www.statmt.org/europarl/
[3]http://www.nilc.icmc.usp.br/lacioweb/

Table 5.1: Results of the translation experiments across three languages

Corpora	Training	Dev.	Test	BLEU	BLEU-N	BLEU-T
Simplext	741	94	90	10.05	9.17	12.56
PorSimplex	741	94	90	48.06	57.97	59.68
Wikipedia	741	94	90	51.43	57.47	58.93

BLEU denotes BLEU scores on the test set; BLEU-N denotes BLEU scores on the test set when no simplification is performed; while BLEU-T denotes BLEU scores on the training data.

three languages, in order to make the three experiments as comparable as possible. Yet again, it was not possible to use the Portuguese part of the Europarl corpora for building the LM (which would make the all three LMs trained on the same domain), as it belongs to the different regional language variety from the one present in the *PorSimples* dataset (Brazilian Portuguese). Therefore, we opted for the Lácio-Web corpus written in the same regional variety as the training dataset, although it does not belong to the same domain as the Europarl texts.

In order to compare the results of translation experiments across the first three corpora (Simplext, PorSimples, and Wikipedia), the systems were trained on the same amount of data. Therefore, we randomly selected only a subset of 925 sentence pairs from the total of 4,483 used by Specia (2010), and a subset of 925 sentence pairs from the total of 137,000 used by Coster and Kauchak (2011a) for the first set of translation experiments.[4]

[4]We first performed translation experiments in English and Brazilian Portuguese on the whole datasets to ensure that our experimental setup led to results comparable with those reported by Specia (2010) and Coster and Kauchak (2011a).

5.3.1 Results of the Automatic Evaluation

The results of the three translation experiments and the sizes of the datasets used are presented in Table 5.1. The results obtained for the PorSimples and Wikipedia datasets achieved reasonable performance (in terms of the BLEU score) of both models in spite of the significantly reduced sizes of the datasets. The same does not hold true for the Spanish dataset, however.

It is interesting to note that for both the PorSimples and Wikipedia datasets, the original sentences compared to the reference simplifications achieve higher BLEU scores than the automatic simplifications compared to the reference simplifications (while this is not the case on the Simplext dataset). Those results indicate that the systems trained on the PorSimples and Wikipedia datasets would achieve very high BLEU scores even if they do not perform any simplifications. This is a consequence of an important difference between cross-lingual MT and monolingual MT. Leaving some segments of the original sentences untranslated does not necessarily deteriorate the grammaticality/fluency of the output in the case of the monolingual MT used for TS, while it seriously deteriorates the output of cross-lingual MT. In the case of MT used for TS, fewer translations performed (or no translations performed at all) does not necessarily mean inadequate output, only less simplified output. Too many substitutions/translations (if inadequate) might lead to a less grammatical/fluent output and a change of the original meaning.

5.3.2 Error Analysis

In order to clarify doubts raised by the results of the automatic evaluation, we performed error analysis on all 90 sentences from the test datasets in English and Spanish. The manual examination of the output of the translation model trained on the Simplext corpora confirmed the poor performance of the system, describing the output of the automatic simplification as barely legible in most of the cases. At the same time, the manual examination of the output of the model trained on the Wikipedia corpus revealed that such a high performance (in terms of the BLEU score) is indeed due to the fact that the output was in most cases identical to the original sentence. The human evaluation of the automatically generated sentences in English and Spanish in terms of their grammaticality, simplicity, and meaning preservation was not possible due to the following reasons:

1. Out of the 90 automatically generated/simplified sentences in Spanish, four sentences were identical to their originals, four differed from their originals only by one word, while all other sentences (82 sentences) were severely ungrammatical (due to the multiple wrong word substitutions), to the extent that it was impossible to assess their simplicity (as they were barely comprehensible) or to compare their meaning with the meaning of the original sentence.

2. Out of the 90 automatically generated/simplified sentences in English, 55 sentences were identical to their originals, 28 sentences differed from their originals by only one word, two sentences differed from their originals by only sentence splitting, two sentences differed from their originals by two words, and three sentences by sen-

tence splitting and one word substitution.

Table 5.2: Examples of automatic simplification in Spanish

Ex.	Original	Automatic	Reference
s1	La **UE** quiere **sacar de la pobreza** a **20** millones de personas **hasta 2020**.	La **Europa** quiere **objetivo** a millones de personas **a**.	Europa quiere ayudar a millones de personas a dejar de ser pobres.
s2	Alex de la Iglesia **dimitirá como presidente** de la Academia de Cine.	**Cine** Alex de la Iglesia **deja la presidencia** de la Academia de Cine.	Alex de la Iglesia dimitirá como presidente de la Academia de Cine.
s3	**Por otro lado**, el informe de **"la Caixa"** sitúa en el 9,5% **del PIB** el déficit público al término de 2010 y **calcula** que el **conjunto** de las **administraciones** podrán reducirlo hasta el 6,4**%** **durante** 2011.	, el informe de **la** sitúa en el 9,5 **mitad el** el déficit público al término de 2010 y **cree** que el **literaria** de las **corridas** podrán reducirlo hasta el 6,4 **mitad en** 2011.	El informe también mostraba que una parte de la deuda en 2010 era de las administraciones públicas. El informe calcula que las administraciones públicas podrán reducir su deuda en 2011.
s4	De este **modo**, ayudaremos a la gente a mejorar su vida, en vez de que **tengan** que empezar desde **cero** después de cada tragedia", añadió.	De este **son peligrosos**, ayudaremos a la gente a mejorar su vida, en vez de que **algunos** que empezar desde **afirma** después de cada tragedia.	Sin las sequías, la vida de las personas mejorará.

The column *Original* contains the original version of the sentence from the test dataset; the column *Automatic* contains the output of the PB-SMT system trained on the Simplext corpus; and the column *Reference* contains the corresponding manually simplified sentence. The differences between the original sentences (*Original*) and the automatic simplification (*Automatic*) are shown in bold.

Table 5.2 shows examples of the original sentences from the test dataset (*Original*), their automatic simplifications (*Automatic*), and their corresponding reference simplifications ('gold standards') manually simplified under the Simplext project (*Reference*). As previously mentioned, 82 out of 90 automatically generated sentences differed from the originals in more than one word, usually encompassing several (multi)word substitutions and deletions.

In the first example (*s1*), "UE" (*EU*) was correctly replaced with "Europa" (*Europe*), while the incorrect substitution of "sacar de la pobreza" (*get out of poverty*) with "objetivo" (*goal/aim/objective*) left the sentence meaningless. Together with the deletion of "20" (in *"20 million people"*) and "hasta 2020" (*until 2020*), and the insertion of "a" at the end of the sentence, the generated sentence is completely ungrammatical and meaningless. The original sentence *"The EU wants to get out of poverty 20 million people until 2020"* is simplified as *"The Europe wants goal to millions of people"*.

The second example (*s2*) is particularly interesting as the manual simplification ('gold standard') is identical to the original sentence. In the automatically generated sentence, however, the phrase "dimitirá como presidente" (*will quit as a president*) in the original sentence was correctly 'translated' as "deja la presidencia" (*leaves the presidency*). One could argue that the phrase used in the automatically simplified sentence is actually simpler than the corresponding phrase in the 'gold standard' (and the original), as the verb "dejar" (*to leave*) is more frequent than the verb "dimitir" (*to quit*). This complies with the common practice in text simplification to replace the infrequent and more specific terms/phrases with their more frequent synonyms. The native speakers might argue that use of the verb "dejar" (*to leave*) introduces ambiguity (as it is not clear whether Alex leaves his presidency because his mandate is over or because he is quitting), while the use of the verb "dimitir" (*quit*) does not leave any doubt about the way/reason Alex is leaving his presidency. Still, non-native speakers will definitely be familiar with the Spanish word "dejar", while (depending on their level of Spanish) may not be familiar with the Spanish word "dimitir".

The third example (*s3*) represents one of the most frequently observed cases of auto-

matic simplification in the test dataset. In those cases, the PB-SMT system generates the output which is at the same time ungrammatical (mostly due to the incorrect deletions of various sentence parts) and meaningless (mostly due to the incorrect word substitutions, but also due to the ungrammatical sentence constructions). For instance, the word "conjunto" (*set*) is replaced with the word "literaria" (*literary*), and the word "administraciones" (*administrations*) with the word "corridas" (*runs*). In the first case, the original word was replaced with the word with a different part-of-speech (a noun replaced with an adjective). However, this example (*s3*) also shows a particularly interesting case of lexical simplification performed by the PB-SMT system, but not performed by the human editor. The word "calcula" (*calculates*) is replaced with the word "cree" (*believes*). In this sentence, the word "calcula" (*calculates*) was indeed used with the meaning "cree" (*believes*), which is not its most common meaning. Such replacements are favourable in text simplification, as stated in Web Content Accessibility Guidelines (W3C, 2008).

In the fourth example (*s4*) only three incorrect substitutions are made. However, as all three words were replaced by words with different parts of speech than the original, not only was the meaning of the sentence changed, but the generated sentence has also become grammatically incorrect. The word "modo" (*way*) in the phrase "de este modo" (*in this way*) was substituted by "son peligrosos" (*are dangerous*) leading to a meaningless and grammatically incorrect phrase "de este son peligrosos" (*in this are dangerous*); the word "tengan" (*have to*) was substituted by the word "algunos" (*some*); and the word "cero" (*zero*) was substituted by the word "afirma" (*affirm/assert*).

Table 5.3 contains several examples of the original sentences from the test dataset

Table 5.3: Examples of the automatic simplification in English

Ex.	Original	Automatic	Reference
e1	Parennes is a commune in the Sarthe department in the region of Pays-de-la-Loire in north-western France.	Parennes is a commune**. It is found** in the Sarthe department in the region Pays de la Loire in north-western France.	Parennes is a commune. It is found in the region Pays de la Loire in the Sarthe department in the west of France.
e2	Mated **females** will lay a clutch of 30-60 eggs in a mound of dry leaves or peat.	Mated **women** will lay a clutch of 30-60 eggs in a mound of dry leaves or peat.	Mated females will lay a clutch of 30-60 eggs in a mound of dry leaves or peat.
e3	Elections to the new **authorities** were held on three different Thursdays in 1973.	Elections to the new **government** were held on three different Thursdays in 1973.	Elections were held to the new authorities on three different Thursdays in 1973.
e4	Whitney Adela Sloan is a British-born actress best known for playing her role as "Hollywood Henderson" in the Disney **Channel** Original **Movie** Go Figure.	Whitney Adela Sloan is a British-born actress best known for playing her role as "Hollywood Henderson" in the Disney **rté** Original **subsidiary** Go Figure.	Whitney Adela Sloan is an English actress.

The column *Original* contains the original version of the sentence from the test dataset; the column *Automatic* contains the output of the PB-SMT system trained on the Wikipedia corpus; and the column *Reference* contains the corresponding sentence from the Simple English Wikipedia. The differences between the original sentences (*Original*) and the automatic simplification (*Automatic*) are shown in bold.

(*Original*), their automatic simplifications (*Automatic*), and their corresponding reference simplifications ('gold standards') from the Simple English Wikipedia (*Reference*). They illustrate some of the phenomena revealed during the manual error analysis.

Example *e1* presents one of the five correctly performed sentence splittings learned by the PB-SMT system. However, it is important to mention that all five split sentences in the test dataset share the same structure of the original sentence ('*X is a commune in...*'). In all five cases, such an original sentence is transformed into two sentences which again share the same structure ('*X is a commune. It is found in...*'). The example *e2* presents an example of a bad word substitution (lexical simplification which leads

to a simpler sentence but changes the original meaning), while *e3* shows a good word substitution (lexical simplification). The example *e4* contains an example of a sentence with two wrong word substitutions. It can be noted that all examples of the automatically simplified sentences are still grammatical. One or two wrongly applied word substitutions may only change the meaning of the sentence but they do not deteriorate the grammaticality of the sentence. Correctly applied word substitutions and sentence splittings preserve the original meaning and grammaticality of the sentence, and lead to a slightly simpler output.

5.4 Sentence Similarity Assessment

The manual analysis of the output generated by the PB-SMT systems trained on the Wikipedia and Simplext corpora revealed significant differences in the simplification strategies. Those differences led to a grammatically correct output with preserved original meaning (in most cases) in the system trained on the Wikipedia corpus, and to a completely ungrammatical output with severely changed meaning (or no meaning at all) in the other system trained on the Simplext corpus. Given that the sizes of the training and development datasets were equal in both systems, and the language models were built using the same corpora in two languages, we discarded the size of the dataset and the type of the sentences used for building the language model as the main factors of the differences in system performance. The results are even more surprising when we take into account the fact that the Simplext training data is obtained from a parallel corpus with a controlled quality of simplifications, and the Wikipedia corpus is built from only comparable data with no quality check (neither in terms of the automatic sen-

tence alignment, nor the simplification quality). Therefore, all four available datasets (the three corpora used in the previously described translation experiments, and the EncBrit corpus) were closely examined in terms of the type of transformation present in them. The four corpora were compared on the basis of ten sentence similarity metrics. The metrics included the cosine similarity, S-BLEU, METEOR (Denkowski and Lavie, 2011), TERp (Snover et al., 2009), and the six sub-metrics of the TERp: the number of insertions (*Ins*), deletions (*Del*), substitutions (*Sub*), shifts (*Shift*), word shifts (*WdSh*), and errors (*Err*).[5]

5.4.1 Sentence Similarity Metrics

Cosine similarity (cosine) uses the bag-of-words representation and is calculated according to the following formula:

$$\text{COSINE}(O, S) = \frac{|O \cap S|}{\sqrt{|O| \times |S|}} \tag{5.1}$$

where O and S represent the bag-of-words in the original sentence (O) and its corresponding simplified version (S).

Sentence-level BLEU score (S-BLEU) differs from BLEU (Papineni et al., 2002) only in the sense that S-BLEU will still positively score segments that do not have higher n-gram matching (n=4 in our setting) unless there is no unigram match; otherwise it is the same as BLEU.

METEOR is designed as a robust, sentence-level metric, which addresses several weaknesses of the BLEU metric: the lack of recall, the use of higher order n-grams

[5] All sentence similarity metrics (except the cosine similarity) were calculated using the mteval-v13a software, available at: http://www.itl.nist.gov/iad/mig/tests/mt/2009/.

for fluency and grammaticality, and the use of geometric averaging of n-grams (Lavie and Denkowski, 2009). The word alignment in METEOR does not only rely on exact matching, but also on stem matching (two words are matched if they share identical stems) and synonymy (words are matched if they are synonyms of each other). When comparing the METEOR results among the four datasets, one should bear in mind that the results between two different languages cannot be directly compared due to the fact that English, Spanish, and Portuguese do not support all types of matching techniques in the latest version of the METEOR (version 1.4) which was used here. While all three languages support the exact match and the stem match, English also supports both synonym and paraphrase match, Spanish supports only the synonym match, and Portuguese does not support either of the two.[6]

TERp is an automatic evaluation metric for MT, which measures the number of 'edits' needed to transform the MT output (simplified version of the original sentence in our case) into the reference translation (original sentence in our case). TERp is an extension of TER – Translation Edit Rate (Snover et al., 2006) that utilizes phrasal substitutions (using automatically generated paraphrases), stemming, synonyms, relaxed shifting constraints and other improvements (Snover et al., 2009). The higher the value of TERp (and each of its components), the less similar the original and its corresponding simplified sentence are. The other three metrics (cosine, S-BLEU, and METEOR) measure the similarity between the original sentence and its corresponding simplified version, i.e. the higher their value, the more similar the sentences.

[6]http://www.cs.cmu.edu/ alavie/METEOR/README.html

Table 5.4: Sentence similarity metrics on the training datasets and EncBrit

Metric	Simplext	Wikipedia	PorSimples	EncBrit
Cosine	**0.32 ± 0.23**	0.78 ± 0.23	0.78 ± 0.19	**0.45 ± 0.17**
S-BLEU	**0.16 ± 0.23**	0.58 ± 0.36	0.58 ± 0.26	**0.15 ± 0.16**
METEOR	**0.20 ± 0.23**	0.75 ± 0.26	0.69 ± 0.23	**0.39 ± 0.18**
TERp	**131.15 ± 94.06**	56.43 ± 94.36	41.14 ± 51.92	**139.93 ± 97.64**
Ins	**13.72 ± 13.44**	4.47 ± 8.91	3.23 ± 5.51	**12.21 ± 11.94**
Del	****1.11 ± 3.16**	1.73 ± 6.68	2.32 ± 3.86	****1.43 ± 3.29**
Sub	**10.21 ± 8.15**	2.56 ± 4.05	3.78 ± 4.10	**6.33 ± 4.34**
Shift	**2.50 ± 2.41**	0.70 ± 1.38	1.03 ± 1.43	**1.80 ± 1.66**
WdSh	**3.44 ± 3.57**	1.02 ± 2.12	2.57 ± 4.26	***2.69 ± 2.68**
Err	**27.54 ± 16.91**	9.45 ± 13.19	10.36 ± 9.07	**21.77 ± 11.51**

Metrics which achieved a significantly different score (at a 0.001 level of significance measured by the independent-samples t-test in SPSS) for Simplext and EncBrit datasets than for both the Wikipedia and PorSimples datasets are shown in bold; those which significantly differ only from the results obtained using the Wikipedia dataset are shown in bold and with one asterisk '*', while those which significantly differ only from the results obtained using the PorSimples dataset are shown in bold and with two asterisks '**'.

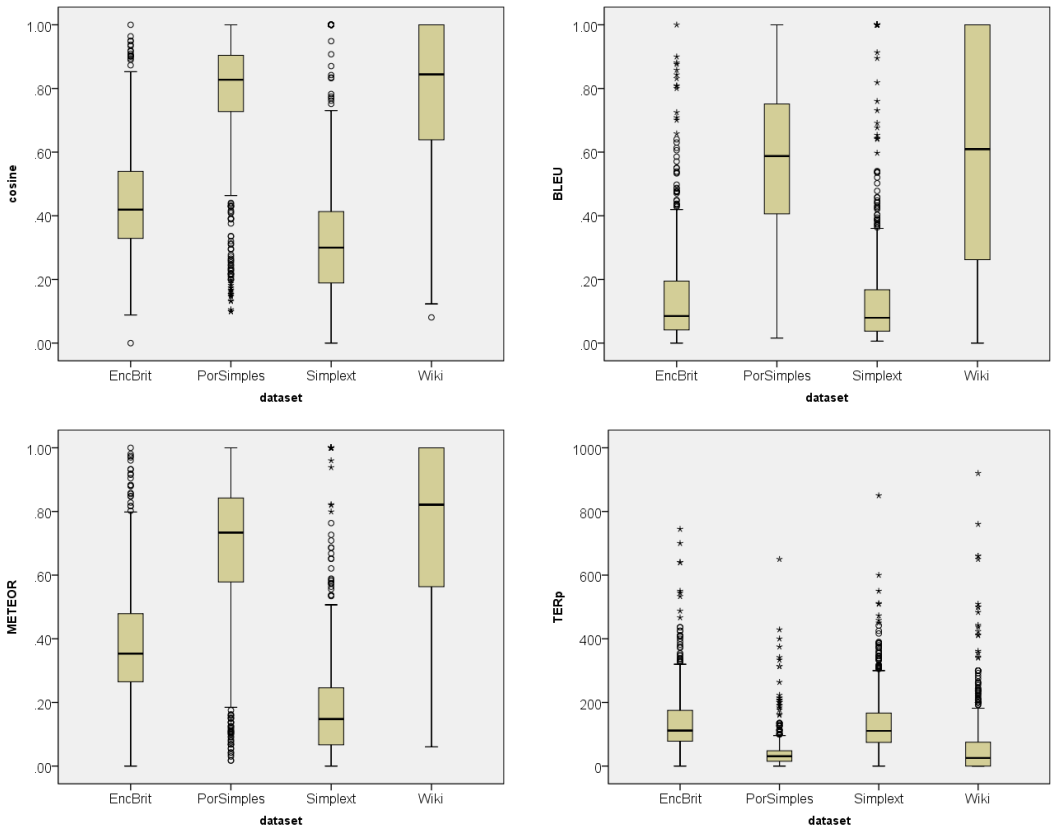
5.4.2 Sentence Similarity Results

The results of the sentence similarity experiment are given in Table 5.4, presenting the mean value and standard deviation of each metric on each dataset. It appears that the similarity between the original and simplified sentences used for training is much higher (up to four times higher in the case of the S-BLEU score) in Wikipedia and PorSimples datasets than in the other two datasets (Simplext and EncBrit). Closer examination of the results obtained for TERp and its components shows that the main difference between the Wikipedia and PorSimples datasets on one side, and the Simplext and EncBrit datasets on the other side, does not lie in the number of deletions but rather in the number of insertions and substitutions.

The distribution of cosine similarity, S-BLEU, METEOR, and TERp across the four

corpora (Figure 5.1) and a closer examination of the S-BLEU scores (Figure 5.2) indicate that the cause of the good performance of the 'translation' system trained on PorSimples and Wikipedia probably lies in the nature of the data.

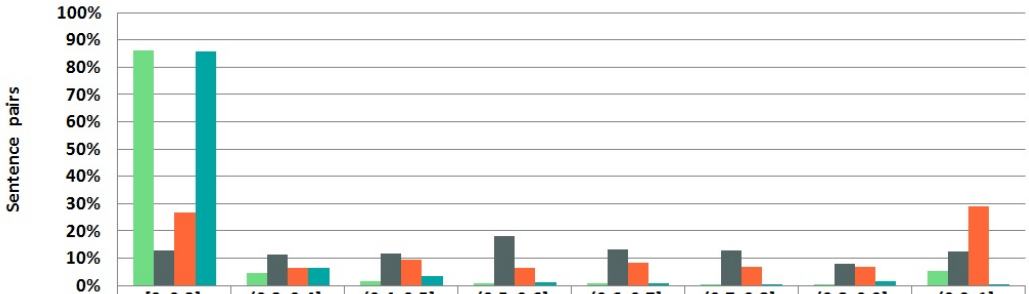
Figure 5.1: Cosine similarity, S-BLEU, METEOR, and TERp across the four datasets



The height of the rectangle indicates the spread of the metric on each corpus, the horizontal line inside the rectangle indicates the mean, while the whiskers outside the rectangle indicate the smallest and largest observations which are not outliers. Outliers are presented with small circles beyond the whiskers.

The Wikipedia corpus contains only those sentence pairs whose normalised similarity was higher than 0.5 (Coster and Kauchak, 2011b). The PorSimples corpus consists only of the sentence pairs simplified by 'natural' simplification in which the most com-

Figure 5.2: Distribution of the S-BLEU scores across the four datasets



	[0, 0.3]	(0.3, 0.4]	(0.4, 0.5]	(0.5, 0.6]	(0.6, 0.7]	(0.7, 0.8]	(0.8, 0.9]	(0.9, 1]
Simplext	85.96%	4.45%	1.62%	0.94%	0.94%	0.27%	0.40%	5.40%
Porsimples	12.96%	11.20%	11.74%	18.08%	13.23%	12.82%	7.83%	12.28%
Wikipedia	26.86%	6.48%	9.31%	6.34%	8.37%	6.88%	6.75%	29.15%
EncBrit	85.86%	6.49%	3.49%	1.16%	0.83%	0.50%	1.50%	0.17%

The percentage of sentences in each dataset with the S-BLEU score in a specific interval; the columns represent the intervals, e.g. *85.96%* in the column *[0, 0.3]* and row *Simplext* means that 85.96% of the sentence pairs in the Simplext corpus have the S-BLEU score in the interval [0, 0.3]).

mon simplifying operation is sentence splitting (Gasperin et al., 2009). EncBrit and Simplext corpora, on the other hand, contain a great number of deletions and strong paraphrases (combinations of lexical and syntactic transformations with deletions) as reported by Bautista et al. (2011), and Štajner et al. (2013). Some of these differences are illustrated in Table 5.5, which contains several examples of sentence pairs with various S-BLEU scores for each of the four corpora. It can be noted that those sentence pairs with a very low S-BLEU score (*w1*, *e1*, *s1*, and *p1*) are very strong paraphrases, sometimes not even preserving the original meaning, and thus cannot represent good training material for the standard PB-SMT system.

Table 5.5: Examples of sentence pairs with various S-BLEU scores

Ex.	S-BLEU	Original	Simple
(w1)	0.02	*Travis's style is well explained and exemplified by Marcel Dadi on the DVD The Guitar of Merle Travis, which includes live video performances by Travis of classics such as "John Henry" and "Nine Pound Hammer" as well as transcriptions of Travis solos in tablature.*	*Travis wrote many well-known songs including : "Sixteen Tons ". Travis is best known for his masterful guitar playing style.*
(s1)	0.02	*Por otra parte, en el tercer trimestre del 2010 se calificaron provisionalmente (planes estatales y autonómicos) 12.188 viviendas protegidas, un 25,8% menos que en el mismo trimestre del año anterior y un 25,6% menos que en el trimestre precedente.*	*La construcción de viviendas protegidas también disminuyó en los últimos meses de 2010.*
(w2)	0.50	*Knowles rose to fame in the late 1990s* as the lead singer of *the* R&B girl group Destiny's Child.	*She was famous first* as the lead singer of R&B girl group Destiny's Child.
(e2)	0.49	*In gold, marble, carved wood, and rare tiles, these interiors are decorated in Baroque, Rococo, or rocaille.*	*The interiors of several of Lisbon's churches are decorated in gold, marble, carved wood, and rare tiles.*
(s2)	0.50	*Licenciada en Bellas Artes por la Universidad Politécnica de Valencia, Ana Juan es ilustradora, escritora y pintora.*	*Ana Juan es ilustradora, escritora y pintora. Estudió Bellas Artes en la universidad de Valencia.*
(p2)	0.50	*Segundo* a campanha*, porém,* os Reis Magos são "os *legítimos* representantes da magia do Natal"*, e não* faltam críticas ao velho *barrigudo.*	*Mas* a campanha *diz que* os Reis Magos são "os *verdadeiros* representantes da magia do Natal". *Não* faltam críticas ao velho *de barriga grande.*
(w3)	0.70	*It also occurs* as a vein mineral in deposits from hot springs, and *it occurs* in caverns as stalactites and stalagmites.	*One can find it* as a vein mineral in deposits from hot springs, and in caverns as stalactites and stalagmites.
(e3)	0.71	Vienna is the *undisputed* cultural centre of Austria and one of the world capitals of music.	Vienna is the cultural center of Austria and one of the world capitals of music.
(s3)	0.69	*Descubierto* un taller ilegal de armas en Alicante.	*La Policía descubre* un taller ilegal de armas en Alicante.
(p3)	0.69	Estamos *cadastrando* ferros-velhos, mas não tivemos um trabalho forte contra a clonagem *e agora* vamos atacar para valer – admitiu Bacci.	Estamos *reunindo informações sobre* ferros-velhos, mas não tivemos um trabalho forte contra a clonagem. *Agora,* vamos atacar para valer – admitiu Bacci.

Examples of sentence pairs with various S-BLEU scores from Wikipedia (*w*), EncBrit (*e*), Simplext (*s*), and PorSimples (*p*) corpora (differences between the original and simple version are shown in italics)

5.5 Quality vs. Quantity

Based on the results of the initial translation experiments (Section 5.3) and sentence similarity experiments (Section 5.4.2), we formulate the following two hypotheses:

- **H1**: The size of the training and development datasets does not significantly influence the performance of the standard PB-SMT model for automatic text simplification.

- **H2**: The type of the sentence pairs in the training and development datasets (in terms of their S-BLEU score) significantly impacts the performance of the standard PB-SMT model for automatic text simplification. We expect that:

  - The sentence pairs with S-BLEU scores below 0.3 do not represent good training material, as such transformations cannot be expected to be learnt with a PB-SMT model. Having a large number of such sentence pairs in the training set could lead to a situation in which the model learns 'bad' simplifications. This would result in output which is worse (less comprehensible and not simpler) than the original.

  - Having a large number of sentence pairs with an S-BLEU score higher than 0.9 would lead to a model which tends to leave the original sentence unchanged. In those sentence pairs, the original sentence and its corresponding simpler version are almost exactly the same. This would not deteriorate the scores for grammaticality and meaning preservation, but would not score high in terms of simplification.

– The sentence pairs with S-BLEU scores between 0.3 and 0.9 represent good training material, and would lead to a slightly simpler output which is grammatical and preserves the original meaning.

Several examples of sentence pairs with various S-BLEU scores from the Wikipedia corpus are presented in Table 5.6.

5.5.1 Translation Experiments using the Wikipedia Corpus

In order to test our hypotheses (H1 and H2), we trained a series of translation models on datasets of varying size and similarity of sentence pairs. All experiments were done only for the English language, as the corpora in the other two languages (Spanish and Brazilian Portuguese) were not large enough. All experiments employed the same standard PB-SMT model described in Section 5.2.2. The corpus of 60,000 Simple English Wikipedia articles (version 2.0 document-aligned data)[7] was used for the language model.[8] The initial dataset of 167,689 aligned sentences from English Wikipedia and Simple English Wikipedia (version 2.0 sentence-aligned data) was tokenised and randomised. Using the simplified sentences as references and the original sentences as translation hypotheses, each sentence pair was ranked by its S-BLEU score and the sentence pairs were categorised into eight different sets based on those scores (Table 5.7).

The experiments were conducted using the 11,030 sentence pairs of each category, with the only exception being the category (0.5, 0.6] in which there was not enough data

[7] http://www.cs.middlebury.edu/ dkauchak/simplification/

[8] In the previous experiments (Section 5.3), the translation results were compared across three languages. As the goal was to make the results as comparable as possible, and there are no large available corpora of simple texts to be used for the language model in Spanish and Brazilian Portuguese, the Europarl corpus was used for the language model in English as well (instead of Simple English Wikipedia which is used here).

Table 5.6: Examples of sentences pairs with various S-BLEU scores from Wikipedia

Ex.	S-BLEU	Original sentence	Simpler version
(w1)	0.03	"*The crown-of-thorns starfish has gained notoriety as a threat to the coral reef ecosystem, particularly in the Great Barrier Reef off the coast of Australia. Overpopulation of crown-of-thorns has been blamed for widespread reef destruction.*"	"*However, when there are too many Crown-of-thorns, they can devastate a coral reef.*"
(w2)	0.08	"*In women, the larger* mammary glands *within* the breast *produce the milk*."	"The breast *contains* mammary glands."
(w3)	0.38	"*Built as a double-track railroad bridge, it* was completed on January 1, 1889, and *went out of service* on May 8, 1974."	"*It was built for trains and* was completed on January 1, 1889. *It closed down* on May 8, 1974 *after a bad fire*."
(w4)	0.47	"*However*, Arizona *still can expect experiencing* the effects of tropical cyclones once every five years*, in average*."	"*In average*, Arizona experiences the effects of tropical cyclones once every five years."
(w5)	0.55	"In 2000, the series *sold its naming rights to* Internet search engine Northern Light *for five seasons, and t*he series was named the Indy Racing Northern Light Series."	"In 2000, the series *sponsor became the* Internet search engine Northern Light. *T*he series was named the Indy Racing Northern Light Series."
(w6)	0.63	"*Wildlife which eat acorns as* an important part of their diet*s* include birds, such as jays, pigeons, some ducks, and several species of woodpeckers."	"*Creatures that make acorns* an important part of their diet include birds, such as jays, pigeons, some ducks and several species of woodpeckers."
(w7)	0.77	"It was *discovered* by Brett J. Gladman in 2000 , and given the *temporary* designation S2000 S 5."	"It was *found* by Brett J. Gladman in 2000, and given the designation S2000 S 5."
(w8)	0.87	"Austen was not well known in Russia *and t*he first Russian translation of an Austen novel did not appear until 1967."	"Austen was not well known in Russia. *T*he first Russian translation of an Austen novel did not appear until 1967."

Differences between the two versions are shown in italics.

(Table 5.7). In order to investigate the impact that the size of the training set has on the quality of the output, we created subsets of the previously categorised data in order to train 40 different models (Table 5.8). For each category of sentences (specific S-BLEU range), we built five translation models using training and development sets of different

Table 5.7: Distribution of S-BLEU in the Wikipedia corpus

S-BLEU	Used in our experiments	Total available	
		Number	Percentage
[0, 0.3]	11,030	42,106	25.11%
(0.3, 0.4]	11,030	13,142	7.84%
(0.4, 0.5]	11,030	11,749	7.01%
(0.5, 0.6]	10,979	10,979	6.55%
(0.6, 0.7]	11,030	11,195	6.68%
(0.7, 0.8]	11,030	11,863	7.07%
(0.8, 0.9]	11,030	11,951	7.13%
(0.9, 1]	11,030	54,692	32.62%

sizes. The five translation models which used 2,000, 4,000, 6,000, 8,000 and 10,000 sentence pairs for training, were tuned on 200, 400, 600, 800, and 1,000 sentence pairs, respectively (Table 5.8).

Table 5.8: The forty PB-SMT systems built in the experiments

S-BLEU	Size of the training set + development set (number of sentence pairs)				
	2,000 + 200	4,000 + 400	6,000 + 600	8,000 + 800	10,000 + 1,000
[0, 0.3]	S-03-200	S-03-400	S-03-600	S-03-800	S-03-1000
(0.3, 0.4]	S-04-200	S-04-400	S-04-600	S-04-800	S-04-1000
(0.4, 0.5]	S-05-200	S-05-400	S-05-600	S-05-800	S-05-1000
(0.5, 0.6]	S-06-200	S-06-400	S-06-600	S-06-800	S-06-1000
(0.6, 0.7]	S-07-200	S-07-400	S-07-600	S-07-800	S-07-1000
(0.7, 0.8]	S-08-200	S-08-400	S-08-600	S-08-800	S-08-1000
(0.8, 0.9]	S-09-200	S-09-400	S-09-600	S-09-800	S-09-1000
(0.9, 1]	S-10-200	S-10-400	S-10-600	S-10-800	S-10-1000

Motivated by previous experience in automatic evaluation of the models using the BLEU score (Section 5.3.1) where the obtained BLEU score heavily depended on the similarity between the original sentence and reference (manual) simplification in the test

set, we tested the 40 new translation models (Table 5.8) on two different test sets:

- The **Wikipedia test set** containing a total of 240 sentence pairs, with 30 sentence pairs from each of the eight categories for the S-BLEU scores ([0,0.3], (0.3,0.4], ... , (0.9,1]);
- The **EncBrit test set** containing all 601 sentence pairs present in the EncBrit corpus.

The first test dataset (*Wikipedia test set*) contains equal numbers of sentence pairs from each of the eight categories ([0,0.3], (0.3,0.4], ... , (0.9,1]) in order to enable a fair comparison of the systems' performances (each system was trained and tuned on the sentence pairs which belong only to one of the eight categories). The second test dataset (*EncBrit test set*) contains an unbalanced number of sentence pairs from each of the eight categories (Figure 5.2, Section 5.4.2).

5.5.2 Results of the Automatic Evaluation

The results of all 40 experiments trained on the Wikipedia corpus and tested on both test datasets (the Wikipedia test set, and the EncBrit test set), varied by the size and the sentence similarity in the training and development datasets, are presented in Table 5.9. It is worth noting that the baseline which does not perform any simplification (leaves the original sentence as it is) achieves a BLEU score of 62.27 on the Wikipedia test set, and 14.51 on the EncBrit test set. Those BLEU scores are calculated using the simplified sentences (from Simple English Wikipedia) as reference translations, and their corresponding original sentences (from English Wikipedia) as translation hypotheses.

Table 5.9: Results of the translation experiments trained on the Wikipedia corpus

S-BLEU	Size of the training set + development set (number of sentence pairs)				
	2,000 + 200	4,000 + 400	6,000 + 600	8,000 + 800	10,000 + 1,000
[0, 0.3]	56.38 (13.84)	56.38 (13.84)	56.15 (13.87)	57.75 (13.68)	57.89 (13.59)
(0.3, 0.4]	60.89 (14.05)	61.35 (13.95)	61.76 (14.08)	61.52 (14.06)	61.37 (14.01)
(0.4, 0.5]	61.27 (14.02)	61.36 (14.09)	61.74 (14.17)	61.55 (14.15)	**62.11** (14.12)
(0.5, 0.6]	60.96 (14.09)	61.30 (14.22)	61.52 (14.27)	61.77 (14.16)	61.98 (14.13)
(0.6, 0.7]	60.96 (14.25)	61.30 (14.30)	61.60 (14.35)	61.69 (14.35)	61.80 (14.32)
(0.7, 0.8]	61.56 (14.30)	61.38 (14.29)	61.67 (14.30)	61.77 (14.30)	61.89 (14.28)
(0.8, 0.9]	61.54 (14.38)	61.49 (14.40)	61.51 (14.40)	61.57 (14.40)	61.61 **(14.41)**
(0.9, 1]	61.57 (12.71)	61.57 (12.52)	61.59 (12.46)	61.55 (12.39)	61.55 (12.54)

BLEU scores for all 40 experiments controlling for the S-BLEU scores on the training and development sets (rows), and for the sizes of the training set (columns); the BLEU scores obtained using the Wikipedia test set are presented outside the brackets, while those obtained using the EncBrit test corpus are presented inside the brackets; the highest scores on each test set are shown in bold. The BLEU score for the baseline which does not make any transformation on the original sentences (i.e. the BLEU score for the test set, using the simplified sentences as reference translations and the corresponding original sentences as translation hypotheses) is 62.27 on the Wikipedia test set, and 14.51 on the EncBrit test set.

As shown in Table 5.9, none of the 40 experiments have even reached the baseline. The only results that are significantly lower than the rest (on both test sets) are those obtained for the experiments in which the training and development datasets consist only of the sentence pairs with S-BLEU scores between 0 and 0.3. Significantly lower than the rest (only on the EncBrit test set) are the results obtained for the experiments trained and tuned on the sentence pairs with S-BLEU scores between 0.9 and 1.

A closer look at the results for each test set (Figures 5.3 and 5.4) suggests that the most probable reason for the poor performance of the translation models actually lies in the nature of the data (similarity of the original and simplified sentences in the training and development datasets), and not in its size, thus supporting both our hypotheses (H1 and H2) formulated in Section 5.5. It is interesting to note that the results of the experiments trained only on the sentence pairs with S-BLEU scores between 0.9 and 1 are

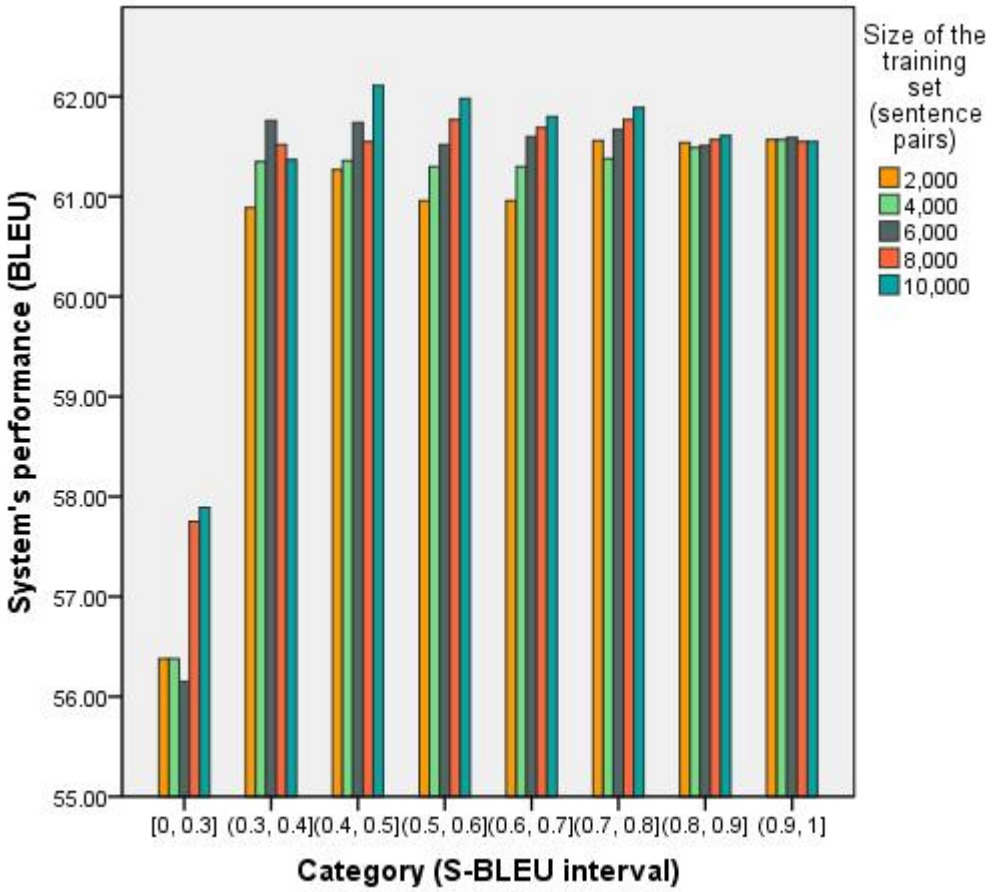


Figure 5.3: System's performances tested on the Wikipedia test set

even worse than for the experiments trained only on the sentence pairs with S-BLEU scores between 0 and 0.3, when the test set comes from a different corpus than the training set (Figure 5.4). This could be seen as a type of over-fitting the data. The results presented in Table 5.9, and Figures 5.3 and 5.4 indicate that the sizes of the training and development datasets do not influence the system's performance significantly for any type of the training or test set used. The only exception to this is the model trained on the sentence pairs with the S-BLEU score in the interval [0, 0.3] tested on the Wikipedia test set where the use of the two larger datasets led to significantly better system perfor-

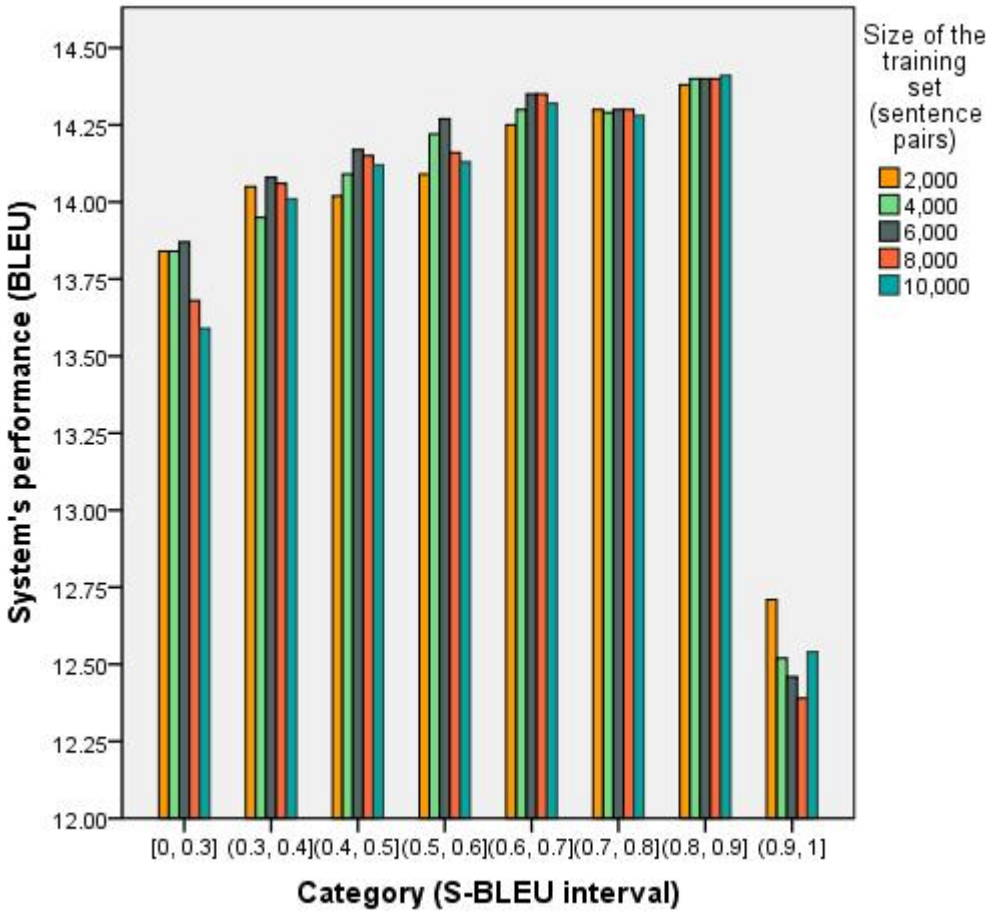


Figure 5.4: System's performances tested on the EncBrit test set

mance (Figure 5.3). The majority of the models tested on the EncBrit test set achieved the best results when trained on 6,000 sentence pairs and tuned on 600 sentence pairs (Figure 5.4).

5.6 Human Evaluation

As we have already shown, the BLEU score on its own does not give a reliable evaluation of the automatically simplified sentences. Therefore, we also conducted a human

assessment of the generated sentences. Following the standard procedure for human evaluation of ATS systems used in previous studies (Coster and Kauchak, 2011a; Drndarevic et al., 2012; Wubben et al., 2012; Feblowitz and Kauchak, 2013; Angrosh and Siddharthan, 2014), three human evaluators were asked to assess the generated sentences on a 1–5 scale (where the higher mark always denotes better output) according to three criteria:

- Grammaticality, i.e. how grammatical the generated output is.
- Simplicity, i.e. how much effort the reader needs in order to read and understand the given sentence.
- Meaning preservation, i.e. how similar the automatically simplified and original sentences are in terms of meaning.

5.6.1 Instructions

Annotators were instructed to try not to penalise the grammaticality of the semantically changed sentences whenever possible. For example, the automatically simplified sentence (8) should get the score 5 for grammaticality, although it does not have a correct meaning (the corresponding original sentence is (9)).

(8) *Mexico lies in an endorheic (characterized by back) drainage basin at an altitude of approximately 7,350 feet (2,240 metres).*

(9) *Mexico lies in an endorheic (characterized by interior drainage) basin at an altitude of approximately 7,350 feet (2,240 metres).*

Table 5.10: Translation systems used in human evaluation

System	Training size	Dev. size	Sentence pairs
S-03-200	2,000	200	$0 <$ S-BLEU < 0.3
S-03-1000	10,000	1,000	$0 <$ S-BLEU < 0.3
S-06-200	2,000	200	$0.5 <$ S-BLEU < 0.6
S-06-1000	10,000	1,000	$0.5 <$ S-BLEU < 0.6
S-10-200	2,000	200	$0.9 <$ S-BLEU < 1
S-10-1000	10,000	1,000	$0.9 <$ S-BLEU < 1

The columns *Training size* and *Dev. size* denote the number of sentence pairs used for the training and development datasets, while the column *Sentence pairs* denotes the category of sentence pairs used in the corresponding systems (in terms of the S-BLEU scores).

To assess simplicity, the annotators were instructed to take into account the choice of words and the syntactic structure of the sentence, and not to penalise the simplicity of the sentence based on grammaticality. For example, sentences (10) and (11) would have the same score '3' for simplicity while in terms of grammaticality and meaning preservation their scores would be different ('5' and '2', respectively).

(10) *It is located in the central Mexican plateau in the Valley of Mexico – more properly a basin – just north of the Neo-Volcanica Range.*

(11) *It is located in the central Mexican plateau, in the Valley of Mexico - more tumors a basin - just north of the Neo-Volcanica park.*

5.6.2 Evaluation Dataset

The main goal for our last set of experiments was to investigate the influence of: (1) the similarity of sentences in the training and development datasets; and (2) the size of the training and development datasets, on the translation performance measured in terms of

grammaticality, simplicity and meaning preservation scores. Therefore, the annotators were asked to rate outputs of different systems (trained on different sizes and categories of the training and development datasets) according to the three aforementioned criteria. In order to obtain statistically sound results, we decided to have at least 20 original sentences in the evaluation dataset. This made the constraint on the number of systems we can evaluate (comparison of all 40 systems would lead to a total of 20*40=800 simplified sentences and 20 original sentences for human evaluation). Therefore, the focus of the human evaluation was only on six out of 40 trained systems (Table 5.10). The categories of the sentence pairs in those six systems correspond to the categories with the highest number of sentence pairs in the previously analysed TS datasets.[9]. The selection of those six systems led to a total of 140 sentences for human evaluation (20 original sentences, and 120 corresponding sentences generated by the selected systems). All sentences were selected from the EncBrit test set. Out of 601 original sentences in the EncBrit test set, we first filtered out all sentences which were left unchanged by at least one of the six systems, following common practice in the human evaluation of ATS systems to only assess the sentences which have undergone at least one modification in each of the systems compared (Feblowitz and Kauchak, 2013; Siddharthan and Angrosh, 2014). The 20 sentences for human evaluation were selected randomly from the remaining original sentences.

[9]Most of the sentence pairs present in the Simplext and Encyclopedia Britannica corpora (85.96% and 85.86%, respectively) belong to the category of sentence pairs with an S-BLEU score between 0 and 0.3; most of the sentence pairs present in the PorSimples corpus (18.08%) belong to the category of sentence pairs with an S-BLEU score between 0.5 and 0.6; while most of the sentence pairs present in the Wikipedia corpus (29.15%) belong to the category of sentence pairs with an S-BLEU score between 0.9 and 1 (Figure 5.2, Section 5.4.2)

5.7 Results of the Human Evaluation

The results of the human evaluation of the six PB-SMT models are presented in Table 5.11. Although the inter-annotator agreement (measured by weighted Cohen's κ) was not high (the average pair-wise agreement among the annotators was 0.59 for grammaticality, 0.44 for simplicity, and 0.46 for meaning preservation), the ranking of the six systems was the same by each annotator (Table 5.11). The system trained on 2,000 sentence pairs with an S-BLEU score between 0.5 and 0.6 (*S-06-200*) achieved the highest score for grammaticality (G) by each annotator (and overall). The best overall score for meaning preservation (M) was achieved by the system trained on 10,000 sentence pairs with the S-BLEU score between 0.5 and 0.6 (*S-06-1000*), although the score for the system trained on a smaller portion of the same category of sentence pairs (*S-06-200*) was not significantly lower. Annotators 2 and 3 rated the system trained on the larger datasets more favourably, while annotator 1 rated the other system (trained on the smaller datasets) better. However, the scores for meaning preservation (M) achieved for those two systems (*S-06-200* and *S-06-1000*) do not differ significantly (measured by the Wilcoxon's sign rank test for repeated measures in SPSS) for any of the three annotators. The system trained on 10,000 sentence pairs with the S-BLEU score between 0 and 0.3 (*S-03-1000*) achieved the best simplicity score by each annotator (and overall).

5.7.1 The Impact of the Size of the Datasets

In order to explore the impact of the size of the datasets on the systems' performance, each pair of experiments (S-03-200 and S-03-1000, S-06-200 and S-06-1000, S-10-200 and S-10-1000) was compared across the three scores (G, M, and S) using the

Table 5.11: Results of human evaluation of the systems

System	Annotator 1			Annotator 2			Annotator 3			All annotators		
	G	M	S	G	M	S	G	M	S	G	M	S
Original	4.90	/	2.50	4.75	/	2.25	4.90	/	2.80	4.85	/	2.60
S-03-200	4.05	3.80	2.30	4.20	4.50	2.55	3.85	3.55	2.85	4.03	3.95	2.57
S-03-1000	4.25	3.80	2.65	4.15	4.50	2.70	4.20	3.80	3.20	4.20	4.03	**2.85**
S-06-200	4.40	4.30	2.55	4.45	4.45	2.55	4.65	4.60	2.95	**4.50**	4.45	2.68
S-06-1000	4.40	4.25	2.50	4.30	4.55	2.55	4.60	4.65	3.10	4.43	**4.48**	2.72
S-10-200	3.25	2.55	2.20	3.60	3.65	2.45	2.90	2.55	2.70	3.25	2.92	2.45
S-10-1000	2.90	2.75	2.15	3.25	3.50	2.55	2.60	2.60	2.90	2.92	2.95	2.53

The column *G* contains the average scores obtained for Grammaticality (called "Fluency" in some of the previous studies); the column *M* contains the average scores obtained for Meaning Preservation (called "Adequacy" in some of the previous studies); and the column *S* contains the average scores obtained for Simplicity of the automatically generated sentences. The highest achieved overall scores (*All annotators*) for G, M, and S (excluding those scores for original sentences) are presented in bold. Annotator 1 is a non-native English speaker at a proficiency level; Annotators 2 and 3 are native English speakers.

Wilcoxon's sign rank test for repeated measures in SPSS. The results suggest that the size of the training and development datasets impacts the system's performance only in two special cases. First, it significantly influences grammaticality of the sentences generated by the PB-SMT models trained and tuned on the sentence pairs with an S-BLEU score between 0.9 and 1. In this case, the influence is the opposite from what was expected; the bigger the datasets, the less grammatical the output is. This is illustrated by the following examples of the original sentence (12), the output of the S-10-200 system (13), and the output of the S-10-1000 system (14), with the deviations from the original sentence shown in bold:

(12) *Mexico lies in an endorheic (characterized by interior drainage) basin at an altitude of approximately 7,350 feet (2,240 metres).*

(13) *Mexico lies in an endorheic (characterized by interior)* **drainage** *basin***,** *at an altitude of approximately 7,350 feet (2,240 metres).*

(14) *Mexico* **is** *an endorheic (characterized by interior)* **drainage** *basin at* **a height can change** *of approximately 7,350 feet (2,240 metres).*

Furthermore, the size of the training and development datasets significantly influences the simplicity of the sentences generated by the PB-SMT models trained and tuned on the sentence pairs with an S-BLEU score between 0 and 0.3; the bigger the datasets, the simpler the output is. The following examples of the output of the S-03-200 system (15), and the output of the S-03-1000 system (16) for the previously mentioned original sentence (12) further illustrate this phenomenon (deviations from the original sentence are shown in bold):

(15) *Mexico lies in an endorheic (characterized by interior)* **drainage** *basin at an altitude of approximately 7,350 feet (2,240 metres).*

(16) *Mexico* **is** *an endorheic (characterized by interior)* **drainage** *basin at an altitude of approximately 7,350 feet (2,240 metres).*

The meaning preservation score does not seem to be significantly influenced by the size of the datasets for any category of sentence pairs used for training and tuning the systems.

The results presented in this section support our hypothesis (H1) that the size of the training and development datasets is not a key factor for the success of the ATS systems built using the standard PB-SMT model.

5.7.2 The Impact of the Sentence Similarity in the Datasets

In order to explore the impact of the level of similarity of sentence pairs used for training and tuning the standard PB-SMT model for ATS on its performance, we compared each pair of experiments trained on the datasets of the same sizes (S-03-200 and S-06-200, S-06-200 and S-10-200, S-03-200 and S-10-200, S-03-1000 and S-06-1000, S-06-1000 and S-10-1000, S-03-1000 and S-10-1000) across the three scores (G, M, and S) using the Wilcoxon's sign rank test for repeated measures. Grammaticality (G) and meaning preservation (M) of the sentences generated by the systems built using only those sentence pairs with an S-BLEU score between 0.5 and 0.6 seem to be significantly better (at a 0.01 level of significance) than of the sentences generated by the systems built only using those sentence pairs with an S-BLEU score between 0 and 0.3. The grammaticality and meaning preservation of the systems trained only on the sentence pairs with an S-BLEU score between 0.9 and 1 are significantly lower (at a 0.01 level of significance) than in any other system. The following example of an original sentence (17), the output of the S-03-1000 system (18), the output of the S-06-1000 system (19), and the output of the S-10-1000 system (20) illustrates this phenomenon (deviations from the original sentence are shown in bold):

(17) *Although largely of postwar construction, this central area retains its old street pattern, and most of the surviving historical and architectural monuments are located there.*

(18) *Although* **mostly** *of postwar construction, this central area retains its old street pattern, and most of the surviving* **and architectural historical** *monuments are*

located there.

(19) *Although* **mostly** *of postwar construction, this central area retains its old street pattern, and most of the surviving historical and architectural monuments are located there.*

(20) **As of the postwar** *construction, in this central area* **uses** *its old street pattern, and most of the* **historical monuments and and architectural** *are located there.*

The simplicity (S) of the generated output seems to be influenced by the level of similarity of sentence pairs only in extreme cases. The simplicity of the sentences generated by the system trained on 10,000 sentence pairs with the S-BLEU score between 0 and 0.3 is significantly better (at a 0.01 level of significance) than the simplicity of the sentences generated by the system trained on 10,000 sentence pairs with the S-BLEU scores between 0.9 and 1. The following original sentence (21), the output of the S-03-1000 system (22), and the ouput of the S-10-1000 system (23) are such an example (deviations from the original sentence are shown in bold):

(21) *Madrid was occupied by French troops during the Napoleonic Wars, and Napoleon's brother Joseph was installed on the throne.*

(22) *Madrid was occupied by French troops during the Napoleonic Wars, and Napoleon's brother Joseph was* **put** *on the throne.*

(23) *Madrid was occupied by French troops during the Napoleonic Wars, and Napoleon's brother Joseph was* **-RRB-** *installed* **on them** *on the throne.*

The results presented in this section support our second hypothesis (H2) that the level of similarity of sentence pairs in the training and development datasets (in terms of their S-BLEU score) significantly impacts the performance of the standard PB-SMT model for ATS. Our results further indicate that the use of the sentence pairs with low S-BLEU scores (between 0 and 0.3) for training and tuning a standard PB-SMT model for ATS lead to a lower grammaticality of the generated sentences. It was surprising to see that the use of the sentence pairs with very high S-BLEU scores (between 0.9 and 1) for training and tuning a standard PB-SMT model for ATS leads to a lower quality of generated sentences in terms of all three measures (grammaticality, meaning preservation, and simplicity). It seems that the sentence pairs with moderate S-BLEU scores (between 0.5 and 0.6) are the most successful in producing grammatical sentences which preserve original meaning. However, sentences generated in that way do not seem to be much simpler than their originals. The results indicate that the sentence pairs with low S-BLEU scores (between 0 and 0.3) are necessary in the training and development datasets in order to generate sentences which are somewhat simpler than their originals. This might be the explanation as to why the ATS system built by Specia (2010) achieves such a good performance in spite of the relatively small size of the training and development datasets. The sentence pairs used for building Specia's ATS system seem to have a much better distribution of S-BLEU scores than the sentence pairs used for building other ATS systems (Figure 5.2, Section 5.4.2).

5.7.3 Comparison with the State-of-the-Art ATS Systems in English

Our experimental setup for the human evaluation follows the previously established standards for this task (Wubben et al., 2012; Feblowitz and Kauchak, 2013; Angrosh and Siddharthan, 2014)[10]. Those three previous studies also evaluate the Simplicity, Fluency (which is here called "Grammaticality"), and Adequacy (which is here called "Meaning Preservation") on a five-point Likert scale. Furthermore, our ATS systems (PB-SMT models) are trained on the same corpus (Wikipedia corpus) as the previously proposed systems (Wubben et al., 2012; Feblowitz and Kauchak, 2013; Angrosh and Siddharthan, 2014). This allows us to, at least roughly, compare our results with the ones obtained for the state-of-the-art ATS in English (Table 5.12).

The grammaticality (G) and meaning preservation (M) of the output of our systems (except those systems trained on the sentence pairs with an S-BLEU score between 0.9 and 1) were rated higher than the output of all previously proposed systems (Table 5.12). However, the simplicity (S) of the sentences generated by all of our systems was rated lower than the simplicity of sentences generated by all previously proposed systems except one of the systems built by Woodsend and Lapata (2011a). These results indicate that the PB-SMT models built on the carefully selected training sets (e.g. sentence pairs with the S-BLEU score in the intervals [0, 0.3] or (0.5, 0.6]) can perform better than the state-of-the-art systems for English in terms of grammaticality and meaning preservation score (the simplicity only outperforms one state-of-the-art system).

It is important to bear in mind that the scores presented in Table 5.12 only allow us to

[10] Narayan and Gardent (2014) use the 0–5 scale instead of the standard 1–5 scale. Their study was thus excluded from this comparison.

Table 5.12: Comparison with the state-of-the-art ATS systems in English

Reference	System	G	M	S
(Wubben et al., 2012)	(Zhu et al., 2010)	2.59	2.82	2.93
	REVH (Woodsend and Lapata, 2011a)	3.18	3.28	2.96
	(Wubben et al., 2012)	3.83	3.71	2.88
(Feblowitz and Kauchak, 2013)	(Feblowitz and Kauchak, 2013)	3.80	3.09	3.55
	(Wubben et al., 2012)	3.64	3.91	3.07
	(Coster and Kauchak, 2011a)	3.74	3.86	3.19
(Angrosh and Siddharthan, 2014)	(Angrosh and Siddharthan, 2014)	3.52	3.40	3.73
	ALIGNED (Woodsend and Lapata, 2011a)	1.97	2.23	2.33
Current study	**S-03-200**	**4.03**	**3.95**	*2.57
	S-03-1000	**4.20**	**4.03**	*2.85
	S-06-200	**4.50**	**4.45**	*2.68
	S-06-1000	**4.43**	**4.48**	*2.72
	S-10-200	*3.25	*2.92	*2.45
	S-10-1000	*2.92	*2.95	*2.53

The column *G* contains the average scores obtained for Grammaticality (called "Fluency" in some of the previous studies); the column *M* contains the average scores obtained for Meaning Preservation (called "Adequacy" in some of the previous studies); and the column *S* contains the average scores obtained for Simplicity of the automatically generated sentences. The scores of our systems which are higher than scores of all previously proposed systems are shown in bold; the scores of our systems which are higher than scores of some (but not all) previously proposed systems are shown with an '*'.

roughly compare the performance of all systems, as the systems evaluated in different studies (column *Reference*) use different test sets for the evaluation. The task of our systems is even more difficult than the task of all other previously proposed systems, as we train them on one corpus (Wikipedia corpus) and test on another corpus (EncBrit corpus). Although both corpora belong to the same genre and domain, the results are still expected to be worse than in the case of training and testing the systems within the same corpus (Wikipedia corpus) which was the case in all previous studies.

5.8 Summary

This chapter presented several sets of experiments which led to a better understanding of a PB-SMT approach to text simplification. The experiments investigated the impact on the system's performance by: (1) the type of the datasets (parallel or comparable); (2) the size of the datasets; and (3) sentence similarity between the original sentences and their corresponding simplified versions in the datasets. The results indicated the following:

1. The type of the datasets (parallel or comparable corpora) does not have any impact on the success of a standard PB-SMT model in text simplification.

2. The size of the training and development datasets does not significantly influence the performance of a standard PB-SMT model for ATS, in general.

   (a) In the extreme case where all sentence pairs in the training and development datasets have an S-BLEU score between 0 and 0.3, more data leads to a simpler output.

   (b) In the extreme case where all sentence pairs in the training and development datasets have an S-BLEU score between 0.9 and 1, more data leads to a less grammatical output.

3. The similarity (in terms of S-BLEU score) of the original sentences and their simplified versions in the training and development datasets significantly influences the quality of the generated output in all three aspects (grammaticality, meaning preservation, and simplicity).

4. BLEU is not a good measure of the performance of a standard PB-SMT model for ATS, as it mainly reflects the similarity between the original sentences and their simplified versions in the test set and not the actual system's performance (due to the important differences between the cross-lingual MT and the monolingual MT used in ATS).

The first finding is very encouraging given that one of the main problems of any data-driven approach to text simplification is the scarcity of the parallel corpora which consists of original sentences and their corresponding manual simplifications. The compilation of comparable TS corpora should be an easier task than the compilation of parallel TS corpora, as it requires less manual work and human expertise.

The second finding rejects the widespread assumption that the success of a PB-SMT approach largely depends on the size of the training and development datasets. The results indicate that the size of the datasets does not significantly influence the system's performance. This is particularly important as one of the main problems in text simplification is not only the scarcity of the parallel copora but also the size of the existing datasets (usually about 1,000 sentence pairs or fewer).

The third finding indicates that the similarity between the original sentences and their corresponding manual simplifications in the training and development datasets has the strongest impact on the performance of the system. This finding can be used to better model the standard PB-SMT models for automatic text simplification by carefully selecting training and development datasets.

Finally, the fourth finding reveals another important difference between cross-lingual MT and the monolingual MT used in TS. While in cross-lingual MT a system which

does not perform any translation/modification achieves a zero BLEU score, in monolingual MT a system which does not perform any translation/modification can achieve any BLEU score (high or low) depending on the test set used. In the monolingual MT, the BLEU score of the system which does not perform any translation/modification on the input sentences is equal to the BLEU score between the original sentences and their reference/manual simplifications in the test set.

CHAPTER 6

EVENTSIMPLIFY: EVENT-BASED ATS SYSTEM

This chapter presents EventSimplify, our automatic text simplification system which simultaneously reduces and simplifies news stories in English. The system employs a semantically motivated, event-based simplification approach built upon a state-of-the-art event extraction system of factual event mentions (Glavaš and Šnajder, 2014). The system discards text which does not belong to any of the extracted event mentions, thus performing significant content reduction. To the best of our knowledge, this is the first such approach to text simplification with the only exception being the recent work of Barlacchi and Tonelli (2013) for Italian. Although based on the same idea of exploiting factual events for text simplification, our task (simplifying news stories) is significantly more complex than that of Barlacchi and Tonelli (simplifying children's stories). Furthermore, we complemented the automatic evaluation with human assessment of grammaticality and information relevance, and offered two different simplification schemes (neither of which was performed by Barlacchi and Tonelli). One of those schemes was further enriched by pronominal anaphora resolution. Additionally, we conducted an in-depth error analysis of the EventSimplify system and compared its performance with that of the state-of-the-art ATS system for English (Woodsend and Lapata, 2011a).

6.1 Motivation

The existing text simplification systems are usually based on sentence splitting as a method for reducing syntactic complexity (e.g. (Siddharthan, 2006; Bott et al., 2012b; Dornescu et al., 2013)). They also add definitions in order to explain unfamiliar concepts and words (Orasan et al., 2013; Saggion et al., 2011). Both those strategies lead to a simplified text which is longer than the original. As already mentioned in Chapter 2, long texts may present an obstacle for certain audiences which have problem to process large amounts of information, such as people with intellectual disabilities (Morgan and Moni, 2008; Gómez, 2011). Therefore, text simplification systems aimed at making texts more accessible to them should not only simplify the written content by using simpler synonyms and splitting long and complex sentences into several simple ones; they should also discard irrelevant information thus performing content reduction which would reduce the memory load necessary for understanding the given text.

The importance of content reduction in text simplification was already emphasised in several studies (Bautista et al., 2011; Saggion et al., 2011). However, to the best of our knowledge, there have been no ATS systems with the full implementation of a content reduction module so far. Our work on detecting sentences in original texts which should be deleted during the simplification process (Chapter 4), as well as the previous work of Drndarević and Saggion (2012), confirmed that this is not a trivial task. Our evaluation of the ATS system proposed under the Simplext project (Drndarević et al., 2013) indicated the lack of a content reduction module as the main reason for the system's performance being far below the human simplification.

Some of the recently proposed data-driven ATS systems perform a certain content reduction (Coster and Kauchak, 2011a; Zhu et al., 2010; Woodsend and Lapata, 2011a). However, the content reduction in those systems is limited to deletion of very short phrases within a sentence, and it is not semantically based (Section 3.2.3, Chapter 3). The absence of semantic knowledge in those systems very often leads to deletion of obligatory arguments in the sentence (Narayan and Gardent, 2014), which results in an output sentence which is ungrammatical and may have a different meaning than the original sentence (Section 3.2.3).

Our EventSimplify system for simplifying news texts written in English addresses all the previously mentioned issues. It performs syntactic simplification with significant content reduction, employing a semantically motivated, event-based simplification approach. The system is motivated by the fact that an event is a dominant information concept in news (Van Dijk, 1985; Pan and Kosicki, 1993). Although news articles typically describe real-world events, the number of descriptive sentences and sentence parts relating to non-essential information in them is still substantial and contributes to the overall complexity of the texts. Such descriptions which do not relate to any of the concrete events and only have the role of providing a wider context to the story, may not be necessary in the context of text simplification. For example, in a news article about the agreement between the Philippine government and China "to diplomatically resolve a tense standoff involving a Philippine warship and two Chinese surveillance vessels in the disputed South China Sea", a fully descriptive sentence such as *"The South China Sea is home to a myriad of competing territorial claims."* may be omitted. News texts also often contain sentences which combine several pieces of information of varying

relevance, as in the following example:

(24) *"Philippines and China diplomatically resolved a tense naval standoff, the most dangerous confrontation between the sides in recent years."*

In the context of TS, it may be desirable to retain only those sentence parts which contain the most relevant information. For example, in the above sentence (24), the *"resolving of a standoff"* is arguably a more relevant piece of information than the *"standoff"* being *"the most dangerous confrontation in years"*.

6.2 Event-Based Text Simplification

A real-world event is a situation that *happens* or *occurs* (Pustejovsky et al., 2003). Representation of a real-world event in a text is a *linguistic event*, usually referred to as an *event mention* (Rosen, 1999). Event mentions consist of *event anchors* and *event arguments*. Event anchors are words that convey the core meaning of an event (e.g. the word *resolved* in example 24). Event arguments are the protagonists and circumstances of events (e.g. *agent*, *time*, *location*).

The core idea behind our simplification approach is to eliminate all elements of the sentence which do not belong to any event mentions. The benefits of a TS system which exploits that idea are two-fold: (1) it reduces text complexity by eliminating irrelevant information; and (2) it increases readability by shortening long sentences. Figure 6.1 illustrates the main idea of the EventSimplify system. Event anchor and event arguments are presented in bold. The *event anchor* is presented in gray, the *agent* in red, the *time* in orange, and the *location* in green.

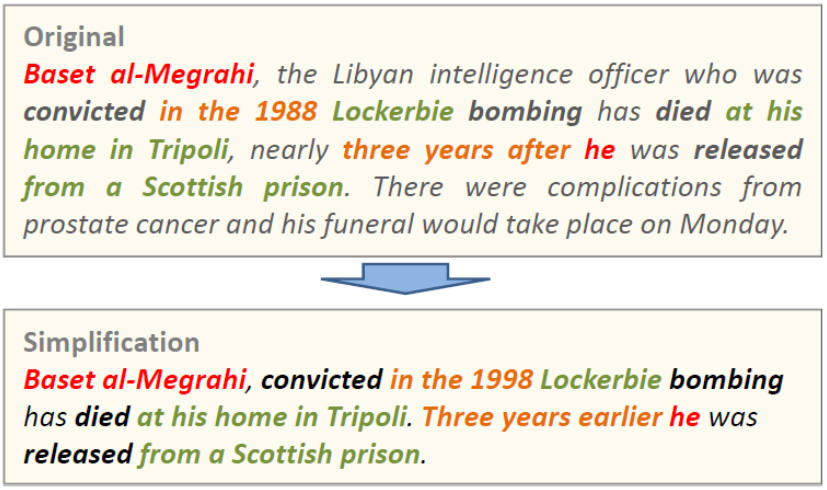


Figure 6.1: Goal of the event-based text simplification

6.2.1 Event extraction system

Our event-based simplification system is built upon a robust event extraction system which involves supervised extraction of factual event anchors (i.e. words that convey the core meaning of the event) and rule-based extraction of event arguments of coarse semantic types (Glavaš and Šnajder, 2014). Given that a thorough description of the event extraction system is outside the scope of this thesis, only the aspects relevant to the proposed simplification schemes will be presented.[1]

For the anchor extraction, the system uses two supervised models, one for identification of event anchors and the other for classification of event type. The first model identifies tokens which are anchors of event mentions (e.g. *"resolved"* and *"standoff"* in *"Philippines and China resolved a tense naval standoff."*), while the second model determines the TimeML event type (Pustejovsky et al., 2003) for previously identified

[1]For more detailed description of the event extraction system see the study by Glavaš and Šnajder (2014).

anchors. The models were trained with logistic regression using lexical and part-of-speech features, syntactic features, and modifier features. The anchor identification model achieves precision of 83%, recall of 77%, and F-score of 80%, while the model for event-type classification performs best for *reporting events*, recognising them with the F-score of 86% (Glavaš and Štajner, 2013).

Event arguments are extracted by a rule-based system which uses a rich set of unlexicalised syntactic patterns on dependency parses (Glavaš and Šnajder, 2013). The system is focused on extracting four coarse-grained argument types: *agent*, *target*, *time*, and *location*, using 13 different extraction patterns in total. Some of those patterns are presented in Figure 6.2 (the argument is shown in bold and the anchor is underlined). The achieved F-score for the argument extraction, evaluated on a held-out set, was: 88.0% for *agent*, 83.1% for *target*, 82.3% for *time*, and 67.5% for *location* (Glavaš and Štajner, 2013).

Name	Example	Dependency relations	Arg. type
Nominal subject	*"**China** confronted Philippines"*	*nsubj(confronted, China)*	Agent
Direct object	*"China disputes **the agreement**"*	dobj(disputes, agreement)	Target
Prepositional object	*"Philippines protested **on Saturday**"*; *"The confrontation **in South China Sea**"*; *"The protest **against China**"*	*prep(protested, on)* and *pobj(on, Saturday)*; *prep(confrontation, in)* and *pobj(in, Sea)*; *prep(protest, against)* and *pobj(against, China)*	Time Location Target
Participial modifier	*"The **vessel** carrying missiles"*; *"The **militant** killed in the attack"*	*partmod(vessel, carrying)*; *partmod(militant, killed)*	Agent Target
Noun compound	*"**Beijing** summit"*; *"**Monday** demonstrations"*; *"**UN** actions"*	*nn(summit, Beijing)*; *nn(demonstrations, Monday)*; *nn(actions, UN)*	Location Time Agent

Figure 6.2: Patterns for argument extraction (Glavaš and Štajner, 2013)

6.2.2 Simplification Schemes

The EventSimplify ATS system offers two different simplification schemes: (1) sentence-wise simplification; and (2) event-wise simplification.

Sentence-wise simplification eliminates all those tokens in the original sentence which do not belong to any of the extracted factual event mentions. This means that only tokens recognised as a part of event anchors or event arguments are preserved in the simplified text. A single sentence of the input text is transformed into a single sentence of the simplified text, assuming it contains at least one factual event mention. Descriptive sentences which do not contain any factual event mentions (e.g. *"Oh what a shame!"*) are eliminated from simplified text. Algorithm 1 summarises the sentence-wise simplification scheme.

Algorithm 1. Sentence-wise simplification

input: sentence s
input: set of event mentions $\mathcal{E}$

```
// initialise the simplified sentence (list of tokens)
```
$\mathcal{S} = \{\}$
```
// list of original sentence tokens
```
$\mathcal{T} = tokenize(s)$
foreach token t **in** $\mathcal{T}$ **do**
  foreach event mention e **in** $\mathcal{E}$ **do**
    `// set of event tokens`
    $\mathcal{A} = anchorAndArgumentTokens(e)$
    `// if the sentence token belongs to an event`
    if t **in** $\mathcal{A}$ **do**
      `// include the token in the simplified sentence`
      $\mathcal{S} = \mathcal{S} \cup t$
      break
output: $\mathcal{S}$

Event-wise simplification transforms each event extracted from the input sentence into a separate sentence of the output. Since a single phrase can be an argument of more than one event mention, a single token from the input sentence may be part of several output sentences. For example, input *"China sent in its fleet and provoked Philippines"* is transformed into output *"China has sent in its fleet. China provoked Philippines"* with *"China"* being the *agent* of both events *"sent"* and *"provoked"*, thus occurring in both output sentences.

In order to retain the grammaticality of the output, we made three additional adjustments to the event-wise simplification:

- Events of the Reporting type (e.g. said) were ignored as they frequently cannot constitute grammatically correct sentences on their own (e.g. "Obama said.").

- Events with nominal anchors were not transformed into separate sentences, as such events tend to have very few arguments, if any. Nominal events are also very often arguments of verbal events. For example, in *"China and Philippines resolved a naval standoff"* the mention *"standoff"* is a *target* of the mention *"resolved"* and has no arguments of its own.

- Gerundive events that govern the clausal complement of the main sentence event were converted into past simple form in the output. For example, the input *"Philippines disputed China's territorial claims, triggering the naval confrontation"* is transformed into *"Philippines disputed China's territorial claims. Philippines triggered the naval confrontation"*, i.e., the gerundive anchor *"triggering"* is transformed into *"triggered"* since it governs the open clausal complement of

the anchor *"disputed"*.

Algorithm 2 summarises the event-wise simplification scheme.

Algorithm 2. Event-wise simplification

input: sentence s
input: set of event mentions $\mathcal{E}$

```
// initialise the set of pairs (event, set of output tokens)
```
$\mathcal{S} = \{\}$
```
// initialise the set of output tokens for each event
```
foreach e **in** $\mathcal{E}$ do
  $\mathcal{S} = \mathcal{S} \cup (e, \{\})$
```
// list of original sentence tokens
```
$\mathcal{T} = tokenize(s)$
foreach token t **in** $\mathcal{T}$ **do**
  foreach event mention e **in** $\mathcal{E}$ **do**
```
    // set of event tokens
```
    $a = anchor(e)$
    $\mathcal{A} = anchorAndArgumentTokens(e)$
```
    // if the token is a part of verbal, non-reporting event
```
    if t **in** $\mathcal{A}$ **&** $PoS(a) \neq N$ **&** $type(t) \neq$ *Rep* **do**
```
     // if the token is a gerundive anchor, it is converted into past simple tense
```
      if $t = a$ **&** $gerund(a)$
        $\mathcal{S}[e] = \mathcal{S}[e] \cup pastSimple(a)$
      else $\mathcal{S}[e] = \mathcal{S}[e] \cup t$
output: $\mathcal{S}$

Additionally, pronominal anaphora resolution was employed on top of the event-wise simplification scheme, as it has been shown that anaphoric mentions cause difficulties for people with cognitive disabilities (Ehrlich et al., 1999; Shapiro and Milkes, 2004). Anaphoric pronouns were resolved using the coreference resolution tool from Stanford Core NLP (Lee et al., 2011).

An example of the original text snippet accompanied by its sentence-wise simplifi-

cation, event-wise simplification, and event-wise simplification with anaphoric pronoun resolution is given in Figure 6.3. Event anchor and event arguments are presented in bold. The *event anchor* is presented in gray, the *agent* in red, the *time* in orange, and the *location* in green. The example in Figure 6.3 also illustrates the imperfections of

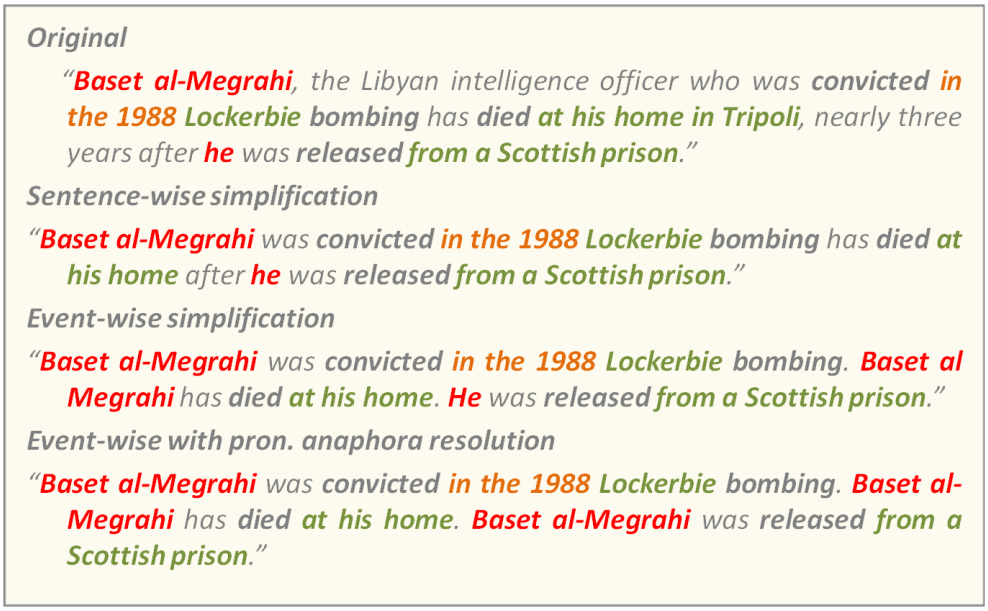


Figure 6.3: An example of event-based text simplification

the current system, which can be addressed in the future. The sentence-wise simplification is not always grammatically correct (*"Baset al-Megrahi was convicted in the 1988 Lockerbie bombing has died at his home..."*), while the event-wise simplification in its current state does not always keep time relations between the sentences (e.g. that Baset al-Megrahi died *after* he was released from a prison).

6.3 Evaluation

The output of the EventSimplify system was evaluated automatically for its readability, and evaluated by humans for its grammaticality, and information relevance. Instead of the commonly used human evaluation of simplicity and meaning preservation, we propose the information relevance score which is more appropriate for ATS systems which perform significant content reduction.

6.3.1 Readability

The readability of the system's output was evaluated on 100 news stories collected from EMM NewsBrief[2]. For each original story and its simplified versions, we computed three frequently used readability scores – Kincaid-Flesch Grade Level (KFGL) (Kincaid et al., 1975), Automated Readability Index (ARI) (Smith and Senter, 1967), and SMOG Index (McLaughlin, 1969), as well as three common-sense indicators of readability: average sentence length (ASL), average document length (ADL), and average number of sentences per document (ANS). As a baseline, we used a syntactically motivated simplification strategy that retains only the main clause of a sentence and discards all subordinate clauses. The main and subordinate clauses were identified using the Stanford constituency parser (Klein and Manning, 2003a).

The results indicate that the event-wise simplification significantly ($p < 0.01$)[3] increases the readability for all measures except the average number of sentences (ANS). Large variation in ANS for event-wise simplification is caused by a large variation

[2] http://emm.newsbrief.eu/NewsBrief/clusteredition/en/latest.html

[3] 2-tailed t-test if both samples are approximately normally distributed; Wilcoxon signed-rank test otherwise

Table 6.1: Readability evaluation (readability formulae)

Original vs.	KFGL	ARI	SMOG
Baseline	-27.70% ± 12.51%	-31.03% ± 12.78%	-13.95% ± 7.93%
Sentence-wise	-30.12% ± 13.93%	-30.73% ± 14.20%	-16.26% ± 9.24%
Event-wise	-50.25% ± 12.59%	-50.89% ± 13.43%	-30.77% ± 10.46%
Pronom. anaphora	-47.76% ± 13.91%	-48.14% ± 14.38%	-29.41% ± 10.56%

Table 6.2: Readability evaluation (common-sense indicators)

Original vs.	ASL	ADL	ANS
Baseline	-38.52% ± 12.13%	-38.52% ± 12.13%	0.00% ± 0.00%
Sentence-wise	-44.34% ± 11.06%	-49.76% ± 11.50%	-9.94% ± 8.72%
Event-wise	-65.48% ± 9.31%	-63.36% ± 12.56%	-9.99% ± 39.70%
Pronom. anaphora	-63.60% ± 10.25%	-61.20% ± 14.37%	-9.99% ± 39.70%

in number of factual events per news story. Descriptive news stories (e.g. political overviews) contain more sentences without any factual events, while sentences from factual stories (e.g. murders, protests) often contain several factual events, forming multiple sentences in the simplified text. Event-wise simplified texts seem to be significantly more readable than sentence-wise simplified texts ($p < 0.01$) in terms of all measures except ANS. Absolute values of the Kincaid-Flesch Grade Level (KFGL), average sentence length (ASL), average document length in words (ADL) and the average number of sentences (ANS) for each simplification scheme are presented in Table 6.3.

6.3.2 Human Evaluation

In line with previous work on text simplification (Knight and Marcu, 2002; Woodsend and Lapata, 2011a; Wubben et al., 2012; Drndarević et al., 2013), grammaticality of

Table 6.3: Absolute values of the readability measures for each simplification scheme

Simplification	KFGL	ASL	ADL	ANS
Original	11.0 ± 3.6	23.8 ± 5.3	315.9 ± 181.6	13.6 ± 8.1
Baseline	7.8 ± 2.0	14.4 ± 3.3	192.1 ± 115.0	13.6 ± 8.1
Sentence-wise	7.5 ± 2.0	13.1 ± 3.3	153.5 ± 84.3	12.1 ± 6.9
Event-wise	5.3 ± 1.4	7.8 ± 1.1	110.1 ± 61.4	14.2 ± 8.3
Pronom. anaphora	5.5 ± 1.5	8.3 ± 1.5	115.7 ± 63.7	14.2 ± 8.3

simplified text was evaluated by human judges. Due to the cognitive effort required for the annotation, the evaluators were asked to compare text snippets (consisting of a single sentence or two adjacent sentences) instead of whole news stories. As a consequence of the differences between our event-based ATS system and the previously proposed ATS systems (Knight and Marcu, 2002; Woodsend and Lapata, 2011a; Wubben et al., 2012; Drndarević et al., 2013), we propose a measure of information relevance (Relevance) – calculated as the harmonic mean of the Relevant Information score (RI) and the Irrelevant Information score (II) – instead of the commonly used scores for simplicity and meaning preservation. The meaning preservation score is defined in a way which penalises any change in the meaning between the original sentence and its corresponding simplification, including any loss of information. Given that the main goal of our ATS system is to eliminate all irrelevant information and to retain and simplify only the relevant information, the loss of irrelevant information is actually desirable and should not be penalised. Therefore, we propose a different kind of human evaluation which is more appropriate for those ATS systems which are expected to – in addition to simplification – perform significant content reduction.

Evaluators were instructed to compare each simplified text snippet with the respec-

tive original, and assign three different scores:

1. Grammaticality score (G);

2. Relevant Information score (RI);

3. Irrelevant Information score (II).

Grammaticality score (G) denotes the grammatical well-formedness of text on a 1–3 scale, where: 1 denotes significant ungrammaticalities (e.g. missing subject or object as in *"Was prevented by the Chinese surveillance craft."*), 2 indicates smaller grammatical inconsistencies (e.g. missing conjunctions or prepositions, as in *"Vessels blocked the arrest Chinese fishermen in disputed waters"*), and 3 indicates grammatical correctness.

Relevant Information score (RI) denotes the degree to which relevant information from the original text is preserved semantically unchanged in the simplified text on a 1–3 scale, where: 1 indicates that the most relevant information has not been preserved in its original meaning (e.g. *"Russians are tiring of Putin"* $\rightarrow$ *"Russians are tiring Putin"*), 2 denotes that relevant information is partially missing from the simplified text (e.g. *"Their daughter has been murdered and another daughter seriously injured."* $\rightarrow$ *"Their daughter has been murdered."*), and 3 means that all relevant information has been fully preserved.

Irrelevant Information score (II) indicates the degree to which irrelevant information has been eliminated from the simplified text on a 1–3 scale, where: 1 means that a lot of irrelevant information has been retained in the simplified text (e.g. *"The president, acting as commander in chief, landed in Afghanistan on Tuesday afternoon for*

an unannounced visit to the war zone."), 2 denotes that some of the irrelevant information has been eliminated, but not all of it (e.g. *"The president landed in Afghanistan on Tuesday afternoon for an unannounced visit.*"), and 3 indicates that only the most relevant information has been retained in the simplified text (e.g. *"The president landed in Afghanistan on Tuesday.*").

A few examples of original sentences and their automatic simplifications produced by the EventSimplify system, together with the assigned human evaluation scores, are presented in Table 6.4. Note that the relevant information score (RI) and the irrelevant information score (II) can, respectively, be interpreted as recall and precision of information relevance. The less relevant information is preserved (i.e. false negatives), the lower the RI score. Similarly, the more irrelevant information is preserved (i.e. false positives), the lower the II score. Considering that the well-performing simplification method should, at the same time, preserve relevant and eliminate irrelevant information, for each simplified text we computed **Relevance score (Relevance)** as the harmonic mean of its relevant information score (RI) and irrelevant information score (II).

The evaluation dataset encompassed 70 original newswire text snippets, each consisting of one or two sentences.[4]. These 70 snippets were simplified using the two proposed simplification schemes (plus the additional scheme with the pronominal anaphora resolution) and the baseline, obtaining in that way four different simplifications per snippet:

1. Baseline;

[4]The dataset is freely available at http://takelab.fer.hr/evsimplify

Table 6.4: Human evaluation examples

Ex.	Original	Simplified	G	RI	II	SM
(a)	*"It is understood the dead girl had been living at her family home, in a neighbouring housing estate, and was visiting her older sister at the time of the shooting."*	*"The dead girl had been living at her family home, in a neighbouring housing estate and was visiting her older sister."*	3	3	3	S
(b)	*"On Facebook, more than 10,000 people signed up to a page announcing an opposition rally for Saturday."*	*"On Facebook, more than 10,000 people signed to a page announcing an opposition rally for Saturday."*	2	3	3	S
(c)	*"Joel Elliott, also 22, of North Road, Brighton, was charged on May 3 with murder. He appeared at Lewes Crown Court on May 8 but did not enter a plea."*	*"Joel Elliott was charged on May 3 with murder. He appeared at Lewes Crown Court on May 8."*	3	2	3	S
(d)	*"For years the former Bosnia Serb army commander Ratko Mladic had evaded capture and was one of the world's most wanted men, but his time on the run finally ended last year when he was arrested near Belgrade."*	*"For years the former Bosnia Serb army commander Ratko Mladic had evaded but his time the run ended last year he was arrested near Belgrade."*	1	2	3	S
(e)	*"Police have examined the scene at a house at William Court in Bellaghy, near Magherafelt for clues to the incident which has stunned the community."*	*"Police have examined the scene at William Court near Magherafelt. The incident has stunned the community."*	3	1	3	P
(f)	*"But opposition parties and international observers said the vote was marred by vote-rigging, including alleged ballot-box stuffing and false voter rolls."*	*"But opposition parties and international observers said ."*	1	1	3	B
(g)	*"Foreign Affairs Secretary Albert del Rosario was seeking a diplomatic solution with Chinese Ambassador Ma Keqing, the TV network said."*	*"Foreign Affairs Secretary Albert del Rosario was seeking a diplomatic solution with Chinese Ambassador Ma Keqing, the TV network said."*	3	3	1	B
(h)	*" On Wednesday, two video journalists working for the state-owned RIA Novosti news agency were briefly detained outside the Election Commission building where Putin was handing in his application to run."*	*"On Wednesday two video journalists were briefly detained outside the Election Commission building. Two video journalists worked for the state-owned RIA Novosti news agency. Putin was handing in his application."*	3	2	2	E

G denotes grammaticality score, *RI* denotes relevant information score, *II* denotes irrelevant information score; while *SM* denotes the simplification method used: *B* – baseline, *S* – sentence-wise, *E* – event-wise, and *P* – pronominal anaphora.

2. Sentence-wise simplification;

3. Event-wise simplification;

4. Pronominal anaphora (event-wise simplification with pronominal anaphora resolution).

This resulted in total of 280 pairs of original and simplified text snippets. The inter-annotator agreement (IAA) was measured on 40 pairs of text snippets independently evaluated by each of the three annotators. Since a moderate agreement was observed[5], the evaluators proceeded by annotating the remaining 240 pairs of text snippets (80 each). Pairwise averaged IAA in terms of three complementary metrics – Weighted Cohen's (κ) coefficient (Cohen, 1968), Pearson's correlation, and Mean Absolute Error (MAE) – is given in Table 6.5.

Table 6.5: IAA for human evaluation

Aspect	Weighted κ	Pearson	MAE
Grammaticality (G)	0.68	0.77	0.18
Relevant Information (RI)	0.53	0.67	0.37
Irrelevant Information (II)	0.54	0.60	0.28

As expected, IAA shows that grammaticality is less susceptible to individual interpretations than information (ir)relevance (i.e. RI and II). Nonetheless, moderate agreement is observed for RI and II as well ($\kappa > 0.5$). Finally, the performance of the proposed simplification schemes on the 70 text snippets was evaluated in terms of

[5]Landis and Koch (1977) describe a moderate agreement as $0.4 < \kappa < 0.6$, whereas $0.6 < \kappa < 0.8$ indicates a substantial agreement.

Grammaticality and Relevance. The results are shown in Table 6.6.

Table 6.6: Grammaticality and Relevance

Scheme	Grammaticality (1–3)	Relevance (1–3)
Baseline	2.57 ± 0.79	1.90 ± 0.64
Sentence-wise	1.98 ± 0.80	2.12 ± 0.61
Event-wise	2.70 ± 0.52	2.30 ± 0.54
Pronominal anaphora	2.68 ± 0.56	2.39 ± 0.57

All the simplification schemes produce text which is significantly more relevant than the baseline simplification ($p < 0.05$ for the sentence-wise scheme; $p < 0.01$ for the event-wise and pronominal anaphora schemes). However, sentence-wise simplification produces text which is significantly less grammatical than the baseline simplification. This is because conjunctions and prepositions are often missing from sentence-wise simplifications as they do not form any event mention. The same issue does not arise in event-wise simplifications where each mention is converted into its own sentence, in which case eliminating conjunctions is grammatically desirable. Event-wise and pronominal anaphora schemes significantly outperform the sentence-wise simplification ($p < 0.01$) in both grammaticality and information relevance. The majority of the mistakes in event-wise simplifications originate from a change of meaning caused by the incorrect extraction of event arguments (e.g. *"Nearly 3,000 soldiers have been killed in Afghanistan since the Talibans were ousted in 2001."* $\rightarrow$ *"Nearly 3,000 soldiers have been killed in Afghanistan in 2001."*).

Overall, the event-wise scheme increases readability and produces grammatical text,

preserving at the same time relevant content and reducing irrelevant content. Combined, experimental results for readability, grammaticality, and information relevance suggest that the proposed event-wise scheme is very suitable for text simplification.

6.4 Manual Analysis of the EventSimplify System

The comparison of the text snippets rated the simplified output very highly in terms of both grammaticality and information relevance. A closer look at the simplified versions of the whole texts, however, raised some important issues in text coherence and the overall performance of the system. An example of original text, its simplification produced by the EventSimplify system (event-wise simplification with pronominal anaphora resolution), and its simplification produced by the state-of-the-art ATS system proposed by Woodsend and Lapata (2011a)[6] are presented in Table 6.7. Here it is important to note that the ATS system built by Woodsend and Lapata (2011a) is not directly comparable with EventSimplify as it additionally performs a lexical simplification which is not covered by our system. The text simplification system proposed by Wubben et al. (2012) reports better results than the system proposed by Woodsend and Lapata (2011a) in terms of grammaticality, meaning preservation and simplicity of the generated output (Table 5.12, Section 5.7.3). However, the system proposed by Wubben et al. (2012) uses a PB-SMT model and thus performs a very limited content reduction. As the main focus of the EventSimplify system is on content reduction, we compare it with the only freely available state-of-the-art ATS system which performs significant content reduction (Woodsend and Lapata, 2011a).

[6] http://homepages.inf.ed.ac.uk/kwoodsen/demos/simplify.html

Table 6.7: Example of the whole text simplification

Original text
"Ex-Soviet leader Mikhail Gorbachev says Russian authorities must annul the parliamentary vote results and hold a new election. Mr. Gorbachev spoke to the Interfax news agency as public indignation was rising over reported election fraud in Sunday's election. The pro-Kremlin United Russia party won less than 50 per cent of the vote, a steep fall from its earlier majority, according to preliminary results. But opposition parties and international observers said the vote was marred by widespread reports of vote-rigging. Thousands of Russians rallied this week in Moscow and St. Petersburg, facing off against police and Interior Ministry troops. Hundreds were beaten and detained."
Simplified by the EventSimplify ATS system (event-wise simplification)
"Hold a new election. Mr. Gorbachev spoke the Interfax news agency. Public indignation was rising over reported election fraud. The pro-Kremlin United Russia party won less than 50 per cent from its earlier majority. The vote was marred by widespread reports. Thousands rallied this week St. Petersburg. Thousands faced Interior Ministry troops. Hundreds were beaten."
Simplified by the state-of-the-art ATS system (Woodsend and Lapata, 2011a)
"Ex-Soviet leader Mikhail Gorbachev says Russian authorities must annul the parliamentary vote results. It hold a new election. Mr. Gorbachev spoke to the Interfax news agency as public indignation was rising over reported election fraud in Sunday's election. The pro-Kremlin United Russia party won less than 50 per cent of the vote, a steep fall from its earlier majority, according to preliminary results. It has opposition parties. But international observers said the vote was marred by widespread reports of vote-rigging. It is Thousands of Russians. Thousands of Russians rallied this week in Moscow and St. Petersburg, facing off against police and Interior Ministry troops. Hundreds were beaten and detained."

Our EventSimplify system produces a significantly shorter output (Table 6.7). This is expected, as it tries to eliminate sentences and sentence parts with irrelevant information. More importantly, our system produces significantly shorter sentences and performs more sentence splitting than the system proposed by Woodsend and Lapata (2011a). This indicates that event-motivated syntactic simplification performs better. It generates output which complies with one of the most important rules in all guidelines for producing easy-to-read texts which states that only one main idea per sentence should be used (Table 2.1, Section 2.3).

In order to better understand the strengths and weaknesses of EventSimplify, we

manually analysed ten texts and their simplified versions (only the event-wise simplification with pronominal anaphora resolution) and compared them with the corresponding simplification by the ATS system proposed by Woodsend and Lapata (2011a). The next two subsections present EventSimplify's pros and cons, in turn.

6.4.1 Correctly Simplified Sentences

Manual analysis revealed two types of original sentences which are consistently simplified correctly by our EventSimplify system:

1. **Reporting sentences** ("X said that [something happened]" → "[Something happened]");

2. **Sentences with multiple actions occurring simultaneously** (when the actions are coordinated with "that", "when", and "and")

The following example of an original sentence (25) and the corresponding output of the EventSimplify system (26) illustrates a simultaneous simplification according to both above-mentioned criteria, as the original sentence reports on multiple actions occurring simultaneously:

(25) *"The Philippine Foreign Affairs Department said that the situation developed Tuesday when the Chinese surveillance ships placed themselves between the Philippine patrol boat BRP Gregorio del Pilar and several Chinese fishing boats, GMA News reported."*

(26) *"The situation developed Tuesday. The Chinese surveillance ships placed themselves the Philippine patrol boat BRP Gregorio del Pilar."*

Although correctly simplifying the reporting sentence and the sentences with multiple actions occurring simultaneously, the output of the EventSimplify system for the above-mentioned original sentence is not perfect (the second simplified sentence is grammatically incorrect and missing important information). However, the same original sentence cannot be correctly simplified with the state-of-the-art ATS system built by Woodsend and Lapata (2011a) either (27):

(27) *"It has the Philippine patrol boat BRP Gregorio del Pilar. The Philippine Foreign Affairs Department said that the situation developed Tuesday when the Chinese surveillance ships placed themselves between many Chinese fishing boats. GMA News reported."*

In fact, the output of the system proposed by Woodsend and Lapata (2011a) changes the original meaning and generates even more ungrammatical sentences.

Another similar example, though more complex, is presented in the following original sentence (28) and its simplified version produced by the EventSimplify system (29):

(28) *"The Chinese Embassy said it had received a report that a dozen Chinese fishing boats had taken refuge in a lagoon of Huangyan Island to escape foul weather when the Philippine gunboat blocked the lagoon entrance and sent 12 Philippine soldiers to harass the Chinese fishermen."*

(29) *"The Chinese Embassy had received a report. A dozen Chinese fishing boats had taken refuge in a lagoon. The Philippine gunboat blocked the lagoon entrance. The Philippine gunboat sent 12 Philippine soldiers. The Philippine gunboat to harass the Chinese fishermen."*

In this example, we can observe two cases of correct entity coreference resolution – *"The Chinese Embassy"* in the first sentence, and *"The Philippine gunboat"* in the fourth sentence of the simplified output. The incorrect fifth sentence of the simplified output is the outcome of the wrong treatment of the complex verb construction "sent [someone] to harass" by the event extraction system, which should not account for two events but rather only one and thus left in the same sentence *"The Philippine gunboat sent 12 Philippine soldiers to harass the Chinese fishermen."*. This error of the event extraction system can, however, be corrected by looking at the output of the original sentence parsed with the Stanford parser, where we find "xcomp(sent-38, harass-43)", which tells us that those two event anchors (*sent*, and *harass*) need to stay in the same sentence after simplification.

The same original sentence (28) simplified by the state-of-the-art ATS system trained on the Wikipedia corpus (Woodsend and Lapata, 2011a) leads to a much more complex output (30). It performs only one (incorrect) lexical substitution (*"had received"* substituted by *"had got"*) and no sentence splitting:

(30) *"The Chinese Embassy said it had got a report that a dozen Chinese fishing boats had taken refuge in a lagoon of Huangyan Island to escape foul weather when the Philippine gunboat blocked the lagoon entrance and sent 12 Philippine soldiers to harass the Chinese fishermen."*

Another good example of reporting sentence simplification with significant content reduction are the following original sentence (31) and the output of the EventSimplify system (32):

(31) *"Playing down the significance of Yitzhak Levanon's trip, the official, who asked not to be identified, said the ambassador went to Egypt on Saturday for farewell meetings with foreign and Egyptian diplomats before his retirement."*

(32) *"The ambassador went to Egypt on Saturday before his retirement."*

The ATS system built by Woodsend and Lapata (2011a) leaves the original sentence (31) unchanged.

6.4.2 Incorrectly Simplified Sentences

The majority of the observed errors in the output of the EventSimplify system were due to the parsing errors and the errors in the event extraction system. Only a few were actually caused by the simplification rules used in the EventSimplify system.

Parsing errors. In the majority of those cases where the generated sentence did not preserve the original meaning, the error occurred due to a parsing error and not due to flaws in the event extraction system or the text simplification system, as in the following example of original sentence (33), and the output of the EventSimplify system (34):

(33) *"Many Egyptians view Israel, which signed a peace treaty with Egypt in 1979 after four wars between the two countries, with hostility."*

(34) *"Many Egyptians view Israel. Israel signed a peace treaty with Egypt in 1979 after four wars with hostility."*

The incorrect attachment of the prepositional phrase *with hostility* to the verb *signed* instead of attaching it to the verb *view* is the result of the incorrect parsing of the original sentence with the Stanford parser (used by the argument extraction module in the event

extraction system). In the parser's output we find "prep_with(signed-7, hostility-24)". Unfortunately, we do not have any power over such parsing errors. If the parser had correctly assigned the prepositional phrase *with hostility* to the verb *view*, the output of our simplification system would have been the following:

(35) *"Many Egyptians view Israel with hostility. Israel signed a peace treaty with Egypt in 1979 after four wars."*

Here it is important to note that the state-of-the-art ATS system built by Woodsend and Lapata (2011a) also performs an incorrect simplification very similar to that of the EventSimplify system:

(36) *"Many Egyptians view Israel. It signed a peace treaty with Egypt in 1979 after four wars between the two countries, with hostility."*

The errors of the event extraction system. The errors in the generated output which would cause a loss of relevant information or change of the meaning were usually caused by imperfections in the argument extraction module of the event extraction system. The following example of an original sentence (37) and its output produced by the EventSimplify system (38) illustrates this phenomenon:

(37) *"The incident followed the killing in August of five Egyptian security guards by Israeli soldiers pursuing militants who had ambushed and killed eight Israelis along the Israeli-Egyptian border."*

(38) *"The incident followed the killed in August by Israeli soldiers. By Israeli soldiers pursued militants. Militants had ambushed and killed eight Israelis along the Israeli-Egyptian border."*

In contrast, the first part of the original sentence (37) was correctly simplified by Woodsend and Lapata's ATS system (39). However, the system produced the ungrammatical second simplified sentence (*"it pursuing"* instead of *"it pursued"*) with loss of information as to who pursued the militants:

(39) *"The incident followed the killing in August of five Egyptian security guards by Israeli soldiers. It pursuing militants who had ambushed and killed eight Israelis along the Israeli-Egyptian border."*

The errors of the simplification rules. The only recurring error caused by the imperfection of the simplification rules used in the EventSimplify ATS system seems to be the **loss of timeline** when simplifying long sentences which report on events that are not mentioned in chronological order. This can be illustrated by the following example of the original sentence (40) and its corresponding simplified version generated by the EventSimplify TS system (41):

(40) *"Israel's ambassador to Cairo has travelled to Egypt for the first time since he and his staff were evacuated from the country in September after protesters stormed the Israeli embassy, a Foreign Ministry official said on Sunday."*

(41) *"His staff were evacuated from the country in September. Protesters stormed the Israeli embassy."*

This is an error caused by our simplification rules. In order to avoid such errors, the EventSimplify system should be modified in such a way that it takes into account subordinating conjunctions (e.g. *since*, *before*, *after*) and their position with regard to the anchors of the event mentions. This would also help position the sentences in the

simplified text in chronological order.

Once again, the same original sentence (40) was even more poorly simplified by the Woodsend and Lapata's ATS system (42):

(42) *"Israel's ambassador to Cairo has traveled to Egypt for the first time since he. A Foreign Ministry official said on Sunday.*

6.5 Summary

In this chapter, we proposed EventSimplify, the first ATS system for English built upon a robust event extraction system. In its current version, the system simultaneously performs syntactic simplification and content reduction by eliminating those sentence parts which do not belong to any event mentions. The system offers two simplification schemes: sentence-wise, and event-wise. Additionally, it performs pronominal anaphora resolution on top of the event-wise simplification scheme. In the event-wise setup, EventSimplify generates one simplified sentence for each event mention present in the original sentence, thus performing sentence splitting where necessary. EventSimplify also presents the only existing ATS system which performs significant content reduction, thus shortening given texts and making them more accessible to those readers who have problem with memory load and who cannot distinguish successfully between relevant and irrelevant information (e.g. people with intellectual disabilities).

The automatic evaluation of EventSimplify confirmed its success in reducing text size and improving its readability (in terms of several automatic readability measures). The human evaluation assessed the output of the EventSimplify system favourably in terms of its grammaticality and information relevance. The in-depth manual analysis of

the whole texts simplified by the EventSimplify system revealed two types of original sentences which are (almost always) simplified correctly, thus leading to a significant reduction of the overall text complexity. The manual analysis also discovered several types of system errors caused by the parser's errors in the underlying event extraction system. Unfortunately, we do not have control over those parsing errors. Still, those errors seem to be present in other ATS systems as well. Most importantly, the manual analysis revealed one recurring error in the system's output caused by the simplification rules in EventSimplify. Those errors can be avoided in future by slightly modifying the simplification rules.

CHAPTER 7

READABILITY INDICES FOR AUTOMATIC EVALUATION OF TS SYSTEMS

With the emergence of automatic text simplification (ATS) systems, the question we are faced with is how to automatically evaluate their performance given that access to the target users might be difficult. The experiments presented in this chapter address this issue. Their goal is to investigate whether some of the already existing readability formulae have a good correlation with the possible obstacles to reading comprehension and thus can be used for automatic evaluation of complexity reduction achieved by text simplification systems. The experiments were conducted on both English and Spanish, enabling the comparison of the most important findings between the two languages. The potential of readability indices in text simplification was further highlighted by various examples of their application.

7.1 Motivation

During their long history (since the 1950s), readability formulae have always been regarded as controversial, triggering endless debates about whether they should be used or not. Many objections have been raised against the earliest readability formulae such as the Flesch-Kincaid Grade Level index (Kincaid et al., 1975) or the Flesch readability score (Flesch, 1949), as they take into account only superficial text features (i.e. sen-

tence and word lengths). Another common criticism of the standard readability indices is that they disagree in their assessment of documents (Kern, 2004). However, DuBay (2004) defends their use, arguing that the important issue is the degree of consistency that each formula offers in its predictions of the difficulty of a range of texts and the correlation of the formulae with reading comprehension test results (see Section 3.4.1 for more details). These two arguments are especially important for the use of readability formulae in automatic evaluation of TS systems where the goal is not to give an absolute measure of text complexity/simplicity but rather compare two versions of the same text. Furthermore, Coleman (1971) and Bormuth (1966) highlighted a close correlation between standard readability metrics and the variables shown to be indicative of reading difficulty. These findings motivated our investigation into the potential correlation between standard readability metrics and the metrics sensitive to the occurrence of linguistic phenomena which present possible obstacles for reading comprehension of various language-impaired readers.

Recent advances in NLP tools and techniques offered the possibility for an automatic computation of more sophisticated readability assessment which takes into account more complex features (e.g. average height of the parse tree, average number of noun and verb phrases, etc.) and gives better readability prediction than the traditional Flesch-Kincaid readability formula (Schwarm and Ostendorf, 2005; Petersen and Ostendorf, 2009). In spite of those findings, the existing ATS systems have still been evaluated by using the old readability formulae based solely on the superficial text characteristics (Woodsend and Lapata, 2011b,a; Zhu et al., 2010), probably due to their simplicity which allows them to be computed automatically with a high precision.

The goal of evaluation of TS systems using readability formulae should not be to determine the exact reading level (complexity) of the simplified texts and thus replace the user-focused evaluation. It should rather enable an easy comparison of:

1. Original and simplified texts in order to assess either the necessary complexity reduction (if comparing original texts with the manually simplified ones); or the achieved complexity reduction (if comparing original texts with the automatically simplified ones);

2. Different text simplification systems (i.e. the level of simplification achieved by different TS systems);

3. Automatically simplified texts with the manually simplified ones (in order to assess whether the automatic simplification achieves the same level of simplification as the manual one);

4. Manually simplified texts with a 'gold standard' (easy-to-read texts which were originally written with the target population in mind) with the aim of assessing whether the manually simplified texts reach the simplicity of the 'gold standard', and thus comply with the easy-to-read standards.

With that goal in mind, it is not necessary that readability formulae give better readability prediction than the complex, cognitively motivated features (reflecting the possible reading obstacles to specific target populations). It would be enough that they correlate well with them. To the best of our knowledge, there have been no previous studies which tried to investigate the suitability of using the existing readability for-

mulae (originally intended for the different purpose of assessing the grade level of text books) for the evaluation of text simplification systems. Therefore, we decided to explore whether some of those widely used readability formulae correlate well with the features which can be regarded as obstacles to reading comprehension to various target populations.

7.2 Methodology

This section describes the methodology employed in order to investigate the suitability of using the existing readability indices for automatic evaluation of text simplification systems. It provides a description of the corpora (Sections 7.2.1 and 7.2.2), readability indices (Sections 7.2.3 and 7.2.4), linguistically motivated features (Section 7.2.5) and the experiments (Section 7.2.6).

7.2.1 Corpora in Spanish

For the experiments in Spanish, four corpora (and their subcorpora) were used (Table 7.1).

FIRST – The FIRST corpus consists of 25 original texts and their corresponding manually simplified versions aimed at people with autism spectrum disorders (ASD)[1], compiled under the FIRST project[2] (Orasan et al., 2013). The texts belong to the news, health, general culture, and literature domains. A more detailed description of the corpus can be found in Chapter 4.

[1] Available at: http://www.first-asd.eu/?q=system/files/FIRST_D7.2_20130228_annex.pdf
[2] http://www.first-asd.eu/

Table 7.1: Characteristics of the corpora in Spanish

Corpus (version)	Genre	Target population	Texts	Sentences	Words
FIRST (original)	Various	All	25	325	7,021
FIRST (simplified)	Various	People with ASD	25	387	6,936
Simplext (original)	News	All	200	1,150	36,545
Simplext (simplified)	News	People with ID	200	1,804	24,154
Automatic (original)	News	All	100	557	18,119
Automatic (rules)	News	People with ID	100	558	18,171
Automatic (syntactic)	News	People with ID	100	656	17,884
Automatic (both)	News	People with ID	100	657	17,938
Noticias Fácil	News	People with ID	200	1,431	12,874

Simplext – The Simplext corpus comprises 200 original news articles (provided by the Spanish news agency Servimedia[3]) and their corresponding manually simplified versions aimed at people with intellectual disability (ID), compiled under the Simplext project[4] (Saggion et al., 2011). A more detailed description of the corpora can be found in Chapter 4.

Automatic – The *Automatic* Simplext corpus consists of 100 original news texts (*original*) and three versions of their corresponding automatically simplified texts, using three different simplification strategies: rule-based lexical transformations (*rules*); a rule-based system for syntactic simplification (*syntactic*); and the combination of both (*both*). Details of those simplification strategies and the corpora can be found in the study by Drndarević et al. (2013). The original articles were obtained from the same source as the Simplext corpus in order to be comparable.

[3] www.servimedia.es
[4] www.simplext.es

NoticiasFácil – The corpus comprises 200 news articles from the Noticias Fácil website[5] written for people with intellectual disability. We compiled this corpus with the aim of having the 'gold standard' for comparison with the manually simplified texts in Simplext, as both corpora share the same domains of the articles.

7.2.2 Corpora in English

The main characteristics of the corpora used for the corresponding experiments for English are given in Table 7.2. In our initial experiments, the goal was to test the hypothesis of the correlation of readability indices with twelve linguistically motivated features which were considered as obstacles to reading comprehension of people with ASD. Given that the FIRST project (which aims at providing a text simplification tool for people with ASD) should cover three different text genres (newswire, healthcare leaflets, and prose/stories/fiction), the focus of our initial study (Štajner et al., 2012) was on exploring whether the correlation between the readability indices and those twelve linguistically motivated features holds irrespective of the text genre. For those text genres, there were no manually simplified texts for people with ASD available which could be used as a 'gold standard'. Therefore, we included the Simple English Wikipedia as a potential 'gold standard'.

News – The News corpus is a collection of reports on court cases in the METER corpus (Gaizauskas et al., 2001) and articles from the Press category of the FLOB (Freiburg-LOB Corpus of British English) corpus.[6] The texts from the FLOB corpus are approximately 2,000 words each. The news articles from the METER corpus are

[5] www.noticiasfacil.es
[6] http://khnt.hit.uib.no/icame/manuals/flob/INDEX.HTM

Table 7.2: Characteristics of the corpora in English

Corpus (version)	Genre	Texts	Sentences	Words
News	Newswire	171	14,556	299,685
Health	Healthcare leaflets	91	6,465	113,269
Fiction	Prose/Fiction	120	18,654	243,655
SimpleWiki	Encyclopaedic	170	17,270	272,445

rather short, none of them containing more than 1,000 words. Therefore, only those texts from the METER corpus which were at least 500 words long were included in this collection.

Health – The Health corpus is a collection of healthcare leaflets for distribution to the general public, contained in categories *A01*, *A0J*, *B1M*, *BN7*, *CJ9*, and *EDB* of the British National Corpus (BNC).[7] Documents in this collection vary in word length.

Fiction – The Fiction corpus is a collection of documents from the Fiction category of the FLOB corpus. Each text in this collection contains approximately 2,000 words.

SimpleWiki – The SimpleWiki contains a random selection of encyclopaedic documents from the Simple English Wikipedia[8]. Each text contains more than 1,000 words. This collection is included as a potential model of accessibility, as the main page of the Simple English Wikipedia (SEW) states that it is for everyone (including children and English language learners). Texts from SEW are supposed to be written using simple English words and grammar. However, the quality of SEW is not checked. Therefore, one of the goals in our initial study (Štajner et al., 2012) was to compare the readability of this "standard" with other types of documents.

[7] http://www.natcorp.ox.ac.uk/
[8] http://simple.wikipedia.org/wiki/Main_Page

As none of the four above-mentioned corpora in English contains corresponding simplified texts, four additional corpora in English (which contain both original texts and their corresponding manually simplified versions) were used to illustrate further possibilities of using readability indices in text simplification. The main characteristics of each of those four additional corpora are presented in Table 7.3.

Table 7.3: Characteristics of the additional corpora in English

Corpus	Aimed at	Version	Code	Texts	SentPerText	WordsPerText
WeeklyReader	Language learners	Advanced	WR-A	100	41.13 ± 15.09	747.72 ± 174.39
		Intermediate	WR-I	100	39.37 ± 13.26	687.52 ± 148.24
		Elementary	WR-E	100	39.38 ± 13.14	621.61 ± 157.26
Enc.Britannica	Children	Original	EB-O	20	27.10 ± 8.91	628.30 ± 198.19
		Simplified	EB-S	20	26.45 ± 9.35	382.35 ± 127.69
Wikipedia	Various	Original	W-O	110	34.55 ± 1.87	716.57 ± 117.82
		Simplified	W-S	110	34.49 ± 1.82	675.07 ± 107.03
En-FIRST	People with ASD	Original	EF-O	25	13.64 ± 3.95	285.68 ± 34.46
		Simplified	EF-S	25	22.92 ± 4.79	311.36 ± 76.82

WeeklyReader – The WeeklyReader corpus comprises 100 texts from *Weekly Reader* and their manual simplifications provided by Macmillan English Campus and Onestopenglish[9] aimed at foreign language learners. The corpus is divided into three sub-corpora – advanced, intermediate and elementary – each representing a different level of simplification. The study by Allen (2009) provides a more detailed description of this corpus.

Enc.Britannica – The Enc.Britannica corpus contains 20 texts from the Encyclopedia Britannica and their manually simplified versions aimed at children – Britannica

[9] http://www.onestopenglish.com/

Elementary (Barzilay and Elhadad, 2003)[10].

Wikipedia – The Wikipedia corpus consists of 110 randomly selected corresponding articles from English Wikipedia (EW) and Simple English Wikipedia (SEW). Here, it is important to note that, in general, articles from SEW do not represent direct simplifications of the articles from EW, they just have a matching topic. For this reason, we did not use complete EW and SEW articles. We only used those sentences in original and simplified versions, which existed in the sentence-aligned parallel corpora version 2.0[11] (Kauchak, 2013).

En-FIRST – The En-FIRST corpus comprises 25 texts on various topics manually simplified for people with ASD, compiled under the FIRST project[12], for the purpose of a piloting task[13]. The texts were simplified by carers of people with ASD in accordance with specified guidelines.

7.2.3 Readability Indices for Spanish

We focused on three readability formulae for Spanish: SSR (Spaulding, 1956), LC (Anula, 2007), and SCI (Anula, 2007). While the first one (SSR) measures 'general' readability of a text, the other two measure more specific types of text complexity, the lexical complexity (LC) and syntactic complexity (SCI). As all three formulae were originally intended for manual computation, we had to make some small modifications in order to enable their automatic computation.

[10] http://www.cs.columbia.edu/ noemie/alignment/
[11] http://www.cs.middlebury.edu/ dkauchak/simplification/
[12] www.first-asd.eu
[13] http://www.first-asd.eu/?q=system/files/FIRST_D7.2_20130228_annex.pdf

The Spaulding's Spanish Readability index (SSR) has already been used for assessing the reading difficulty of fundamental education materials for Latin American adults of limited reading ability and for the evaluation of text passages of foreign language tests (Spaulding, 1956). Therefore, it is reasonable to expect that this formula could be used for estimating the level of simplification performed by text simplification systems aimed at making texts more accessible for the same target population (adults of limited reading ability). The index predicts the relative difficulty of reading material based on the vocabulary and sentence structure, using the following formula:

$$SSR = 1.609 \times \frac{|w|}{|s|} + 331.8 \times \frac{|rw|}{|w|} + 22.0 \tag{7.1}$$

Here, $|w|$ and $|s|$ denote the number of words and sentences in the text, while $|rw|$ denotes the number of rare words in the text. In his original formula, Spaulding (1956) considers as *rare words* those words which cannot be found on the list of 1500 most common Spanish words provided in his study (Spaulding, 1956), plus some special cases of numbers, names of months and days, proper and geographic names, initials, diminutives and augmentatives, etc. The SSR index used in our experiments can be seen as a simplified (slightly modified) version of the original Spaulding's index (Spaulding, 1956). In order to enable a precise and consistent automatic computation, we only considered the words not found on Spaulding's list (Spaulding, 1956) as *rare words*.

The Lexical Complexity index (LC) was suggested by Anula (2007) as a measure of lexical complexity of literary texts aimed at second language learners. It is calculated using the following formula:

$$LC = \frac{LDI + ILFW}{2} \tag{7.2}$$

where *LDI* and *ILFW* represent the *Lexical Density Index* and *Index of Low-Frequency Words*, respectively:

$$LDI = \frac{|dcw|}{|s|}, \tag{7.3}$$

$$ILFW = \frac{|lfw|}{|cw|} \times 100 \tag{7.4}$$

Here, $|dcw|$, $|s|$, $|lfw|$, and $|cw|$ denote the number of distinct content words, sentences, low-frequency words, and content words (nouns, adjectives, verbs, and adverbs), respectively. According to Anula (2007) the *low frequency words* are those words whose frequency rank in the Reference Corpus of Contemporary Spanish (CREA)[14] is lower than 1,000.[15]

The Sentence Complexity Index (SCI) was proposed by Anula (2007) as a measure of sentence complexity in a literary text aimed at second language learners. It is calculated by the following formula:

$$SCI = \frac{ASL + ICS}{2} \tag{7.5}$$

where *ASL* denotes the *Average Sentence Length*, and *ICS* denotes the *Index of Complex Sentences*. They are calculated as follows:

$$ASL = \frac{|w|}{|s|}, \tag{7.6}$$

$$ICS = \frac{|cs|}{|s|} \times 100 \tag{7.7}$$

[14] http://corpus.rae.es/lfrecuencias.html

[15] Both lists (from the Reference Corpus of Contemporary Spanish (CREA) and the Spaulding's list of 1500 most common Spanish words) were lemmatised using Connexor's parser in order to retrieve the frequency of the lemma and not a word form (action carried out manually in the two cited works), and to enable a fully automatic computation of both indices.

Here, $|w|$, $|s|$, and $|cs|$ denote the number of words, sentences and complex sentences in the text, respectively. With the aim of computing the SCI index completely automatically, we considered as complex any sentence which contains multiple finite predicates according to the output of Connexor's Machinese parser[16]. The original definition of a complex sentence used by Anula (2007) relies on a manual detection of complex sentences and thus cannot be used for a precise, fully automatic computation of the index.

7.2.4 Readability Indices for English

Our focus was on four widely used readability indices for English: the Flesch Reading Ease Score (Flesch, 1949), the Flesch-Kincaid Grade Level index (Kincaid et al., 1975), the Fog Index (Gunning, 1952), and SMOG grading (McLaughlin, 1969). All four indices were computed completely automatically, using the GNU *style* package[17].

The Flesch Reading Ease score is calculated according to the following formula:

$$Score = 206.835 - (1.015 \times ASL) - (84.6 \times ASW) \tag{7.8}$$

Here, ASL is the average sentence length and ASW is the average number of syllables per word. The Flesch Reading Ease Formula returns a number from 0 to 100. On this scale, documents with a Flesch Reading Ease score of 30 are considered *very difficult* while those with a score of 70 are considered *easy* to read. Flesch (1949) reported that documents presenting fictional stories lay in the range $70 \leq Score \leq 90$. Only comics were assigned a higher score for reading ease than this. The most difficult type of document was that of scientific literature, with $0 \leq score \leq 30$. During the 1940s, the

[16] www.connexor.eu
[17] https://www.gnu.org/software/diction/

Reading Ease Scores of news articles were at the sixteenth grade level. It is estimated that in contemporary times, this has been reduced to eleventh grade level.

The Flesch-Kincaid Grade Level (FKGL) index is a simplified version of the Flesch Reading Ease score. It is based on identification of the average sentence length (ASL) and the average number of syllables per word (ASW) in the document to be assessed. The formula estimates grade level (GL), according to the following equation:

$$GL = (0.4 \times ASL) + (12 \times ASW) - 15 \tag{7.9}$$

This formula was applied to assess the readability of course materials accessed by Navy technical-training students.

The Fog Index exploits two variables: the average sentence length (ASL) and the number of *hard words* (HW) for each 100 words of a document. *Hard words* are considered all those words which contain more than two syllables. This index returns the Grade Level (GL) of the input document, according to the formula:

$$GL = 0.4 \times (ASL + HW) \tag{7.10}$$

The Fog Index predicts an average reading Grade Level of 10 for news articles (Gunning, 1952).

The SMOG grading is computed by considering the *polysyllable count* (*PSC*), equivalent to the number of words that contain more than two syllables in 30 sentences, and applying the following formula:

$$SMOG = 3 + \sqrt{PSC} \tag{7.11}$$

The SMOG formula is quite widely used, particularly in the preparation of US health-care documents intended for the general public.[18]

7.2.5 Linguistically Motivated Features

We focused on twelve features which can be seen as a means of detecting the occurrence of the different types of obstacles to reading comprehension faced by people with ASD (Table 7.4).[19]

Table 7.4: Linguistically motivated complexity features for experiments in English

#	Code	Feature
1	Verb	Average number of verbs per sentence
2	Adj	Average number of adjectives per sentence
3	Adv	Average number of adverbs per sentence
4	Det	Average number of determiners per sentence
5	Noun	Average number of nouns per sentence
6	Prep	Average number of prepositions per sentence
7	CC	Average number of coordinating conjunctions per sentence
8	CS	Average number of subordinating conjunctions per sentence
9	ASL	Average sentence length (measured in words)
10	AWL	Average word length (measured in characters)
11	Pron	Average number of pronouns per sentence
12	Senses	Average number of word senses per word (using WordNet)

The first eight features represent indicators of structural complexity, features 9 and 10 are common indicators of lexical and syntactic complexity, and the last two features represent indicators of semantic ambiguity (Table 7.5). Our main goal was to investigate whether there is a correlation between those twelve linguistically motivated features

[18]For example, the Harvard School of Public Health provides guidance to its staff on the preparation of documents for access by senior citizens that is based on the SMOG formula (http://www.hsph.harvard.edu/healthliteracy/files/howtosmog.pdf, last accessed 1st March 2012).

[19]More details on reading obstacles for people with ASD can be found in (Štajner et al., 2012, 2014a).

and the Flesch Reading Ease score (Flesch, 1949). If such a correlation exists, then the Flesch Reading Ease score might be suitable to measure the simplicity achieved by TS systems.

Table 7.5: Features as indicators of reading obstacles

Code	Indicator of
Verb	Properties of and relations between concepts/entities
Adj	Descriptive information about concepts/entities
Adv	Descriptive information associated with properties of and relations between concepts/entities
Det	References to concepts that are not proper names, acronyms, or abbreviations
Noun	References to concepts/entities
Prep	Prepositional phrases (a well-cited source of syntactic ambiguity and complexity)
CC	Coordinated phrases
CS	Subordinated phrases, including phrases embedded at multiple levels
ASL	Syntactic complexity
AWL	Lexical complexity
Pron	Anaphoric references
Senses	Semantic ambiguity

For the experiments in Spanish, the average number of senses per word was calculated in two ways, using two different lexical sources – EuroWordNet, and Open Thesaurus (Table 7.6). The Spanish EuroWordNet (Vossen, 1998) contains 50,526 word meanings and 23,370 synsets. The Spanish Open Thesaurus (version 2)[20] contains 21,831 target words (lemmas) and provides a list of word senses for each word. Each word sense is presented as a list of substitute words. The total number of word senses is 44,353. All features were automatically extracted: features 1–11 using the output of the Connexor Machinese parser, and features 12 and 13 using additional lexical resources (WordNet for English, and EuroWordNet and Open Thesaurus for Spanish). For com-

[20]http://openthes-es.berlios.de

Table 7.6: Linguistically motivated complexity features for the experiments in Spanish

#	Code	Feature
1	Verb	Average number of verbs per sentence
2	Adj	Average number of adjectives per sentence
3	Adv	Average number of adverbs per sentence
4	Det	Average number of determiners per sentence
5	Noun	Average number of nouns per sentence
6	Prep	Average number of prepositions per sentence
7	CC	Average number of coordinating conjunctions per sentence
8	CS	Average number of subordinating conjunctions per sentence
9	ASL	Average sentence length (measured in words)
10	AWL	Average word length (measured in characters)
11	Pron	Average number of pronouns per sentence
12	SenseWN	Average number of senses per word (using EuroWordNet)
13	SenseOT	Average number of senses per word (using Open Thesaurus)

putation of features 12 and 13, only the lemmas present in the lexical resources used were considered. All occurrences of such lemmas were considered, including repeated lemmas.

7.2.6 Experiments

After the readability indices and linguistically motivated complexity features were extracted for each text, five sets of experiments were conducted (Table 7.7).

The first set of experiments was conducted in order to select the features (out of the initial 13 features) which could potentially be correlated with the readability indices in Spanish. Given that all 13 features reported significantly different values in the original and the corresponding simplified texts (Table 7.8, Section 7.3), they were all used in the next set of experiments. The second set of experiments indicated many significant correlations between the complexity features and the readability indices, and the third set of experiments reported a high linear correlation between each two readabil-

Table 7.7: Experiments

Set	Experiments	Language	Corpora
I	Comparison of the complexity features and the readability indices between the original and simplified texts	Spanish	Simplext, FIRST
II	Correlation between the complexity features and the readability indices	Spanish English	Simplext, FIRST News, Health, Fiction, SimpleWiki
III	Correlation among the readability indices	Spanish English	Simplext, FIRST News, Health, Fiction, SimpleWiki
IV	Comparison of the average sentence length and the readability indices across the corpora	Spanish English	FIRST, Simplext, Automatic, NoticiasFácil WeeklyReader, Enc.Britannica, Wikipedia, En-FIRST
V	Comparison of the paired relative differences of the readability indices across the corpora	Spanish English	FIRST, Simplext, Automatic WeeklyReader, Enc.Britannica, Wikipedia, En-FIRST

ity indices in English (Section 7.5). The fourth and fifth set of experiments had the aim of presenting the various possibilities of using readability indices in text simplification (Section 7.6). The fourth set of experiments illustrated the possibility of assessing: the necessary complexity reduction (by comparing the original texts with the manually simplified ones in the Simplext, FIRST, WeeklyReader, Enc.Britannica, Wikipedia, and En-FIRST corpora); the complexity reduction achieved (by comparing the original texts with the automatically simplified ones in the Automatic corpora); and the success of the manual simplification in reaching the 'gold standard' (by comparing the manually simplified texts in Simplext with the texts in NoticiasFácil). The fifth set of experiments explored the possibility of using the paired relative differences of the readability indices for comparing different text simplification strategies for English and Spanish.

7.3 Differences between Original and Simplified Texts

In the first set of experiments, 13 linguistically motivated features (Table 7.6) and three readability indices (Section 7.2.3) were compared on the pairs of original and manually simplified texts in two corpora in Spanish – Simplext and FIRST (Section 7.2.1). The goal was to investigate which of those features which detect the occurrence of different types of reading obstacles differ significantly between the two versions of the same texts. Those features can be regarded as a means to measure the linguistic obstacles for reading comprehension of the target populations (people with ASD and people with ID). The results of this set of experiments are presented in Table 7.8.

The results indicate that the main simplification strategies in the Simplext corpus were sentence splitting – reflected in a decrease in coordinating conjunctions (*CC*) – and the elimination of adjectives. In the FIRST corpus, however, the main simplification operations were the removal of prepositional phrases (*Prep*) and adjectives (*Adj*). Although a decrease in prepositions (*Prep*), adjectives (*Adj*), average sentence length (*ASL*), and two lexical complexity indices (*LC* and *SSR*) was present in both corpora, the decrease was more pronounced in the Simplext corpus. These observations draw on some important differences in the simplification performed when having in mind the people with ID (Simplext project), and when having in mind the people with ASD (FIRST project). It appears that the first target population needs a higher level of text simplification, including more sentence splitting (reflected in a decrease in coordinating constructions and verbs) and more elimination of adjective and prepositional phrases (reflected in a greater decrease in adjectives and prepositions than in the FIRST corpus).

Table 7.8: Differences between original and simplified texts

Feature	Simp(O)	Simp(S)	P.R.Diff.	Sign.	FIRST(O)	FIRST(S)	P.R.Diff.	Sign.
Verb	3.46	1.97	**−39.88%**	0.000	3.08	2.92	+0.83%	0.397
Adj	2.41	0.67	**−70.96%**	0.000	1.65	1.32	**−15.13%**	0.003
Adv	0.75	0.46	**−21.11%**	0.000	0.97	0.84	−7.86%	0.157
Det	4.97	2.35	**−50.19%**	0.000	3.19	2.83	−8.77%	0.058
Noun	10.99	4.53	**−57.49%**	0.000	7.07	6.21	**−8.43%**	0.022
Prep	6.61	2.35	**−62.79%**	0.000	3.97	3.08	**−20.28%**	0.000
CC	1.00	0.22	**−74.85%**	0.000	0.90	0.74	−3.33%	0.067
CS	0.63	0.35	**−27.97%**	0.000	0.51	0.50	+4.18%	0.826
ASL	32.87	13.54	**−57.63%**	0.000	23.00	19.90	**−10.52%**	0.012
AWL	5.06	4.81	**−4.79%**	0.000	4.92	4.90	−0.25%	0.596
Pron	1.81	0.73	**−53.71%**	0.000	1.77	1.56	−8.41%	0.108
SenseWN	3.78	4.01	**+6.99%**	0.000	3.98	4.11	+3.68%	0.069
SenseOT	3.52	3.65	**+4.47%**	0.000	3.37	3.47	**+3.10%**	0.006
SCI	54.73	35.95	**−34.42%**	0.000	46.53	45.69	+1.01%	0.699
LC	21.05	12.76	**−39.06%**	0.000	18.53	16.17	**−12.88%**	0.000
SSR	184.20	123.82	**−32.60%**	0.000	149.74	139.61	**−6.69%**	0.002

The columns *Simp(O)*, *Simp(S)*, *FIRST(O)*, and *FIRST(S)*, contain the mean value of the corresponding feature on each subcorpus, where *(O)* denotes the original texts, *(S)* the simplified texts, and *Simp* the Simplext corpus. The columns *P.R.Diff.* and *Sign.* present the mean value of the paired relative differences for the two subcorpora from the antecedent two columns, and the two-tailed statistical significance of the differences measured by the paired t-test and rounded at three decimals. Differences which are statistically significant at a 0.05 level of significance are shown in bold. *P.R.Diff.* are calculated according to equation 7.12 in Section 7.6.1.

It is interesting to note that while the average number of pronouns (*Pron*), which is an indicator of ambiguity in meaning, is lower in simplified than in original texts, the other four features which indicate ambiguity in meaning (*SenseWN* and *SenseOT*) show the opposite trend. This is somewhat surprising as we would expect to find a lower number of senses per word in simplified texts than in their corresponding originals, if we assume that ambiguous words present obstacles for the target population. However, it is a common lexical simplification strategy to replace infrequent words with their more frequent synonyms, and long words with their shorter synonyms. Given that the shorter words are usually more frequent (Balota et al., 2004), and that the frequent words tend

to be more ambiguous than the infrequent ones (Glanzer and Bowles, 1976), this lexical simplification strategy would result in having a greater number of ambiguous words and more senses on average per word in the simplified texts than in their corresponding originals. The justification for those substitution decisions might lie in the previous findings from cognitive psychology that the words with the highest number of possible meanings are actually understood faster, due to their high frequency (Cuetos et al., 1997; Jastrzembski, 1981).

Furthermore, the results presented in Table 7.8 indicate the possibility of finding some correlation between the three readability indices (SCI, LC, and SSR) and the linguistically motivated features. First, all indices show significant differences between original and simplified texts. The only exception to that is the case of the SCI index on the FIRST corpora. This is not surprising because the SCI measures syntactic complexity, and there is no significant difference between the linguistic features which would indicate a possible syntactic simplification (*Verb* and *CC*) in the original and simplified texts of the FIRST corpus. Therefore, the similarity of the SCI values for the original and simplified texts of the FIRST corpus should not be taken as a sign of SCI not being adequate for estimating the level of syntactic simplification in general, but rather as a specificity of the FIRST corpus, simplified for people with ASD. Second, the relative differences of SCI, LC, and SSR between original and simplified texts are higher for the Simplext than for the FIRST corpus. This corresponds to the higher relative differences in the frequencies of linguistically motivated features in the Simplext than in the FIRST corpus, thus indicating a possible correlation between the readability indices and those linguistically motivated features.

7.4 Correlation between Readability Indices and Linguistically Motivated Features in Spanish

Given that all 13 features reported significantly different values for the original and the corresponding simplified texts (on the Simplext corpus), they were all used in the next set of experiments which aimed at investigating whether those features are significantly correlated with the three aforementioned readability indices for Spanish. The results of the second set of experiments are presented in Table 7.9.

Table 7.9: Spearman's correlation between readability indices and linguistically motivated features for texts in Spanish

Features	Simplext			FIRST		
	SCI	LC	SSR	SCI	LC	SSR
Verb	**.867**	**.503**	**.571**	**.774**	.009	−.001
Adj	**.550**	**.540**	**.732**	.143	*.336	**.584**
Adv	**.429**	**.215**	**.259**	**.478**	.061	−.146
Det	**.662**	**.620**	**.621**	**.412**	**.434**	**.476**
Noun	**.585**	**.723**	**.810**	**.338**	**.678**	**833**
Prep	**.658**	**.704**	**.759**	**.398**	**.592**	**.782**
CC	**.543**	**.621**	**.703**	**.411**	**.365**	.237
CS	**.604**	**.163**	**.158**	**.576**	−.088	−.148
ASL	**.751**	**.678**	**.756**	**.593**	**.591**	**.675**
AWL	**.169**	**.326**	**.517**	−.177	−.125	**−.413**
Pron	**.644**	**.567**	**.577**	**.418**	.213	.035
SenseWN	.031	**−.267**	**−.246**	.202	**−.386**	**−.504**
SenseOT	− .017	*−.099	*−.128	.134	−.166	−.049

The first three columns present the results obtained using the Simplext corpus, and the last three the results obtained using the FIRST corpus. Statistically significant correlations (at a 0.001 level of significance) are presented in bold, while those not significant at a 0.001 level but significant at a 0.05 level are presented with an '*'.

It can be noted that the readability indices show a significant correlation with many of the linguistically motivated features in both corpora (Simplext and FIRST). As expected, the readability index which measures syntactic complexity of a text correlates

best with the average number of verbs (*Verb*), while the other two readability indices correlate best with the average number of nouns (*Noun*) in both corpora. This is not surprising for the LC which measures lexical complexity of a given text. In the case of SSR, which is supposed to measure a 'general' complexity of a text, this indicates that the index is more sensitive to the features of lexical than syntactic complexity. It can also be noted that the SSR correlates better than LC with most of the features (for both corpora), the only exceptions being the average number of subordinate conjunctions (*CS*) in the FIRST corpus, and the average number of senses per word (*SenseWN*) in the Simplext corpus, computed using the Spanish EuroWordNet.

The two features indicating semantic ambiguity (*SenseWN* and *SenseOT*) are negatively correlated with the two readability indices which take into account lexical complexity of the text (*LC* and *SSR*). It seems that the higher the average number of senses per word, the less complex the text. This brings us back to the previous discussion (Section 7.3) about more frequent words (which lead to text being perceived as lexically less complex in terms of LC and SSR) being more ambiguous than their less frequent synonyms, but still easier to disambiguate and understand.

Additionally, we investigated the correlation among the three readability indices for Spanish (Table 7.10). It seems that all three indices are mutually correlated for the texts in the Simplext corpus. However, for the FIRST corpus, only LC and SSR seem to be correlated. The mutual correlation of all three indices on *all* texts is probably just a reflection of their mutual correlation on the Simplext corpus, as the number of texts in the Simplext corpus (200) is significantly higher than the number of texts in the FIRST corpus (25). For the Simplext corpus, the correlation seems to be highest between the

Table 7.10: Pearson's correlation among three readability indices for Spanish

Corpora	Index	Spearman's			Pearson's		
		SCI	LC	SSR	SCI	LC	SSR
All	SCI	1	*.435	*.515	1	*.418	*.481
	LC		1	***.739**		1	***.731**
	SSR			1			1
Simplex	SCI	1	*.444	*.533	1	*.427	*.496
	LC		1	***.748**		1	***.740**
	SSR			1			1
FIRST	SCI	1	.204	.242	1	.254	.265
	LC		1	***.686**		1	***.658**
	SSR			1			1

Correlations significant at a 0.01 level of significance are shown with an '*'. The highest correlations between two indices for each corpus are shown in bold.

LC and SSR indices. This indicates that the SSR reflects the lexical complexity more than the syntactic complexity of a given text. In all cases where the correlation exists, the correlation is linear.

7.5 Correlation between Readability Indices and Linguistically Motivated Features in English

In English, we first investigated how well the four readability indices are mutually correlated (Table 7.11). Given that three out of four indices are computed as a linear combination of average sentence length and the average word length (in characters or syllables), it was reasonable to expect that they might be mutually linearly correlated. As it can be observed, all four readability indices (Flesch, FKGL, Fog and SMOG) are mutually (almost perfectly) linearly correlated (at a 0.001 level of significance), both on all texts and on each of the corpora separately.

Given the almost perfect linear correlation among the four readability indices for

Table 7.11: Pearson's correlation among four readability indices for English

Corpora	Index	Flesch	Kincaid	Fog	SMOG
All	Flesch	1	−.959	−.957	−.972
	Kincaid		1	**.987**	.950
	Fog			1	.979
	SMOG				1
News	Flesch	1	−.954	−.954	−.971
	Kincaid		1	**.985**	.932
	Fog			1	.969
	SMOG				1
Health	Flesch	1	−.945	−.915	−.931
	Kincaid		1	.974	.947
	Fog			1	**.986**
	SMOG				1
Fiction	Flesch	1	−.957	−.953	−.973
	Kincaid		1	**.993**	.944
	Fog			1	.965
	SMOG				1
SimpleWiki	Flesch	1	−.936	−.942	−.959
	Kincaid		1	.973	.925
	Fog			1	**.978**
	SMOG				1

All correlations are significant on a 0.01 level of significance. The highest linear correlation between two indices for each corpus is shown in bold. The higher the value of the Flesch index, the easer (less complex) the document is. For other indices, the higher values indicate more complex (more difficult to read) documents.

English, only the results of the Spearman's correlation between the FKGL index and the twelve linguistically motivated features are presented in Table 7.12. All twelve features show a significant correlation with the Flesch-Kincaid Grade Level (FKGL) index for all four corpora (the only exception being the average number of pronouns per sentence in the Health and SimpleWiki corpora).

If we exclude the average sentence length (*ASL*) which is one of its components, the Kincaid readability index seems to have the highest correlation with the average number of prepositions (*Prep*), nouns (*Noun*), adjectives (*Adj*), and determiners (*Det*)

Table 7.12: Spearman's correlation between the Flesch-Kincaid Grade Level (FKGL) index and linguistically motivated features for texts in English

Features	News	Health	Fiction	SimpleWiki
Verb	**.411**	**.291**	**.654**	**.392**
Adj	**.744**	**.747**	**.912**	**.718**
Adv	**.219**	**.294**	**.667**	**.435**
Det	**.776**	**.686**	**.861**	**.656**
Noun	**.781**	**.811**	**.907**	**.435**
Prep	**.818**	**.795**	**.928**	**.743**
CC	**.463**	**.653**	**. 729**	**.584**
CS	**.505**	*.247	**.617**	**.255**
ASL	**.912**	**.880**	**.923**	**.847**
AWL	**.712**	**.544**	**.638**	**.686**
Pron	**−.142**	.068	**.345**	.006
Senses	**−.451**	**−.283**	**−.582**	**−.400**

Each column represents results of the correlation experiments for one of the four corpora (News, Health, Fiction, SimpleWiki). Statistically significant correlations (at a 0.001 level of significance) are presented in bold, while those not significant at a 0.001 level but significant at a 0.05 level are presented with an '*'.

per sentence for the News, Health, and Fiction corpora. In the case of SimpleWiki corpus, the FKGL index still has the highest correlation with the average number of prepositions (*Prep*) and adjectives (*Adj*) per sentence, though not as high as for the other three corpora. The correlation of the Kincaid readability index with the average number of nouns per sentences (*Noun*) is significantly lower for the SimpleWiki corpus than for the other three corpora. The average number of pronouns per sentence (*Pron*) is the only feature that does not show a significant correlation with the FKGL index on all corpora, and in those two corpora where it does, the correlations are of the opposite signs. This indicates that the FKGL index might not be the best measure of text complexity in terms of semantic ambiguity, and more especially anaphoric references. Similar to Spanish, the average number of word senses per word (*Senses*) is negatively correlated with the Kincaid readability index.

7.6 Use of Readability Indices in Text Simplification

As already mentioned in Section 7.1, the goal of the evaluation of TS systems using readability formulae should not be to replace the user-focused evaluation by determining the exact reading level (complexity) of the simplified texts. The goal should be to enable an easy comparison of different versions of the same text.

This section illustrates the possible uses of readability indices in text simplification by: (1) comparing the values of readability indices across various corpora in Spanish and English (Section 7.6.1); (2) comparing the achieved complexity reduction (in terms of readability indices) by different text simplification strategies/systems in Spanish and English (Section 7.6.2); and (3) comparing the achieved complexity reduction (in terms of readability indices) of our automatic text simplification systems for English, among themselves and with various manual simplification strategies (Section 7.6.3).

7.6.1 Comparing Readability Indices across Various Corpora

The first possible use of the readability indices in text simplification was investigated by comparing their values on four different corpora in Spanish (Section 7.2.1). The results of those experiments, comparing the average sentence length (ASL), and the three readability indices (SCI, LC, and SSR) on different corpora, are presented in Table 7.13. Each metric is presented as the mean value with standard deviation.

The results presented in Table 7.13 provide various interesting insights. For example, the comparison of the results obtained for *Simplext (original)* and *Automatic (original)* show that the starting point (original texts) in both manual and automatic text simplification under the Simplext project had similar values of *ASL*, *SCI*, *LC*, and *SSR*

Table 7.13: Comparison of readability indices across the corpora in Spanish

Corpus	ASL	SCI	LC	SSR
FIRST (original)	23.00 ± 5.47	46.53 ± 9.29	18.53 ± 3.18	149.74 ± 25.63
FIRST (simplified)	19.90 ± 5.46	45.69 ± 9.97	16.17 ± 4.09	139.61 ± 27.01
Simplext (original)	32.87 ± 6.34	54.73 ± 10.16	21.05 ± 3.58	184.20 ± 19.10
Simplext (simplified)	13.54 ± 1.97	35.95 ± 12.40	12.76 ± 4.46	123.82 ± 24.13
Automatic (original)	33.43 ± 5.58	56.42 ± 9.37	21.57 ± 3.90	182.21 ± 21.65
Automatic (rules)	33.41 ± 5.61	56.48 ± 9.17	21.28 ± 3.86	174.85 ± 20.97
Automatic (syntactic)	28.10 ± 5.28	49.63 ± 10.19	20.21 ± 3.85	174.40 ± 21.44
Automatic (both)	28.16 ± 5.54	50.01 ± 10.23	19.99 ± 3.66	167.21 ± 20.51
NoticiasFácil	9.26 ± 2.13	30.22 ± 10.88	12.23 ± 4.87	104.50 ± 30.02

(i.e. the original texts were of similar complexity). Therefore, the ideal automatic simplification should result in texts with a similar value in those four features as the manually simplified texts in the *Simplext (simplified)* sub-corpus. Comparison of the results obtained for *Automatic (both)* with those for *Simplext (simplified)* on all four features indicates how far from ideal (achieved by manual simplification) is the performance of the automatic simplification. Furthermore, the comparison of the manually simplified texts in *Simplext (simplified)* with those in *NoticiasFácil* (which can be considered as a 'gold standard' of texts aimed at people with intellectual disabilities) could serve as an additional reference point as to whether the performed manual simplification complies with the standards for easy-to-read texts.

Readability indices were also compared across the four corpora in English (Section 7.2.2). None of those four corpora (*WeeklyReader*, *Enc.Britannica*, *Wikipedia*, and *En-FIRST*) contains automatically simplified texts. However, the manually simplified texts in all four corpora were obtained using different simplification strategies, depending on the specific needs of each target population in mind (language learners, children,

the general public, and people with ASD). Therefore, the comparison of the readability indices across those four corpora provide valuable insights into the results of different simplification strategies.

Table 7.14 presents the average sentence length (ASL), and the four readability indices (Flesch, Kincaid, Fog, and SMOG) across these four corpora for English. Each metric is presented as the mean value with standard deviation. It is worth noting that

Table 7.14: Comparison of readability indices across the four additional corpora (and their sub-corpora) in English

Corpus	ASL	Flesch	FKGL	Fog	SMOG
WR-A	19.14 ± 3.51	64.20 ± 8.86	9.06 ± 1.94	12.14 ± 2.13	10.96 ± 1.47
WR-I	18.23 ± 3.20	67.27 ± 8.29	8.40 ± 1.79	11.43 ± 2.01	10.46 ± 1.42
WR-E	16.33 ± 2.65	72.32 ± 7.80	7.23 ± 1.61	10.19 ± 1.78	9.64 ± 1.30
EB-O	22.39 ± 3.27	53.96 ± 5.48	11.29 ± 1.40	14.63 ± 1.52	12.71 ± 0.96
EB-S	14.93 ± 1.33	67.33 ± 5.12	7.59 ± 0.89	10.43 ± 1.31	10.04 ± 0.95
W-O	19.66 ± 2.82	60.55 ± 8.92	9.70 ± 1.63	12.73 ± 1.91	11.35 ± 1.42
W-S	17.70 ± 2.96	65.74 ± 10.45	8.49 ± 1.94	11.41 ± 2.29	10.45 ± 1.69
EF-O	24.69 ± 6.69	56.76 ± 17.15	11.47 ± 3.61	14.63 ± 3.69	12.12 ± 2.56
EF-S	14.45 ± 3.28	74.99 ± 11.33	6.39 ± 2.08	9.42 ± 2.27	9.12 ± 1.74

Key: *WR-A* – WeeklyReader for advanced language learners; *WR-I* – WeeklyReader for intermediate language learners; *WR-E* – WeeklyReader for elementary language learners; *EB-O* – original versions of Enc.Britannica; *EB-S* – Enc.Britannica's simplified versions for children; *W-O* – texts from the original English Wikipedia; *W-S* – texts from the Simple English Wikipedia; *EF-O* – original texts from the FIRST project; *EF-S* – simplified versions of the texts from the FIRST project, aimed at people with ASD.

the original articles from the Enc.Britannica corpus (*EB-O*) and the original articles from the En-FIRST corpus (*EF-O*) have similar complexity in terms of average sentence length (*ASL*) and the four readability indices (*Flesch*, *FKGL*, *Fog*, and *SMOG*). Their simplified versions (*EB-S* and *EF-S*) have a similar average sentence length, but the overall complexity of the texts simplified for people with ASD (*EF-S*) seems to be one grade level lower than the complexity of the texts simplified for children (*EB-S*).

The articles for advanced language learners (*WR-A*) and the original articles from English Wikipedia (*W-O*) have similar complexity in terms of average sentence length (*ASL*) and the four readability indices (*Flesch*, *FKGL*, *Fog*, and *SMOG*). The corresponding simplified versions of the texts from English Wikipedia (*W-S*) seem to be of a comparable complexity (in terms of readability indices) with the WeeklyReader's version of texts aimed at language learners at the intermediate level (*WR-I*). The complexity of simplified Wikipedia articles (*W-S*) is one grade level higher than the complexity of the texts simplified for children (*EB-S*) or language learners at the elementary level (*WR-E*), and two grade levels higher than the complexity of the texts simplified for people with ASD (*EF-S*).

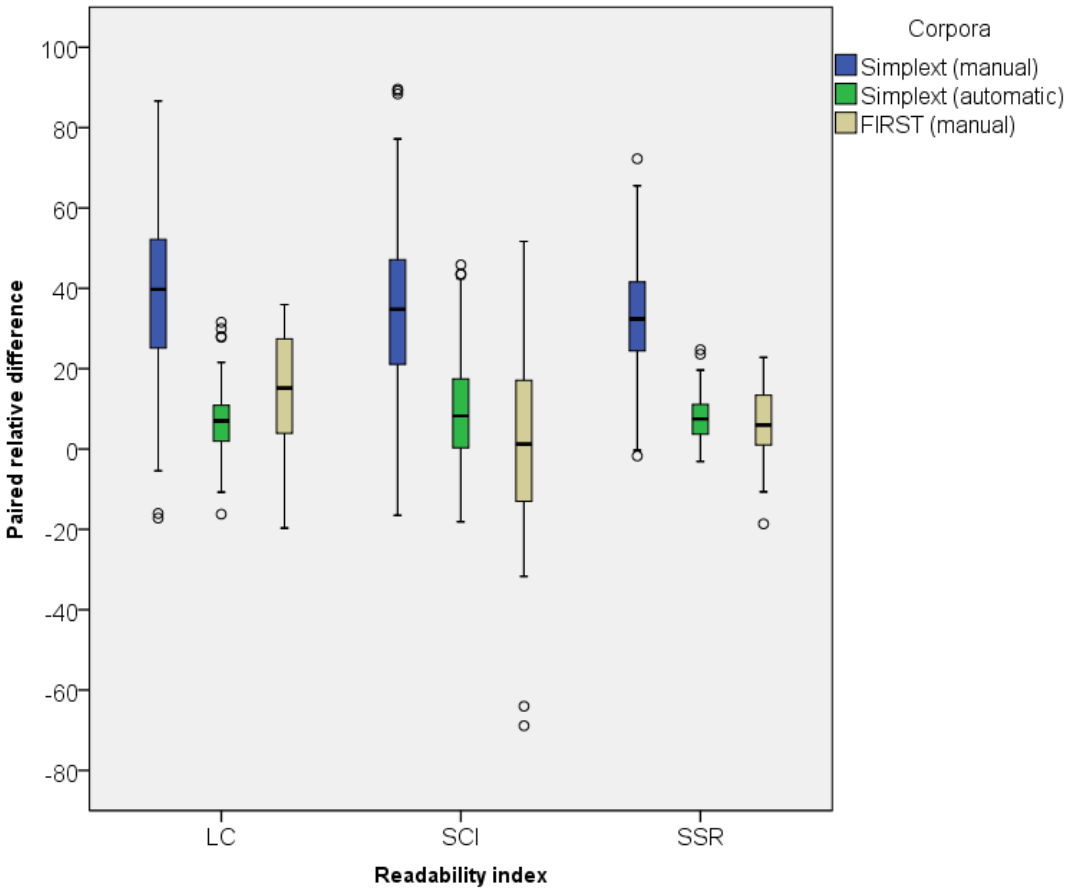
It is also interesting to point out that although there are discrepancies between the grade level assessment between the FKGL index and the Fog and SMOG readability indices, all three indices rank the texts according to their difficulty in the same way. This supports the claims of DuBay (2004) that each formula is consistent in its predictions of the difficulty of a range of texts, although different formulae do not agree in their assessment of the grade level required by each document.

7.6.2 Comparison of Various Text Simplification Strategies

The proposed readability indices can also be used for comparing or ranking of different simplification systems by the level of simplification achieved. Figure 7.1 illustrates this potential of readability indices by providing a quick overview of various text simplification systems for Spanish.

The level of simplification achieved is measured as the mean value of the paired rela-

Figure 7.1: Comparison of different text simplification strategies for Spanish



The y-axis contains paired relative differences (PRD) of each readability index (*LC*, *SCI*, and *SSR*) for the corresponding text pairs in each of the three corpora: *Simplext (manual)*, *Simplext (automatic)*, and *FIRST (manual)*. The height of the rectangle indicates the spread of the PRD on each corpora, the horizontal line inside the rectangle indicates the mean, while the whiskers outside the rectangle indicate the smallest and largest observations which are not outliers. Outliers are presented with small circles beyond the whiskers. If the mean of the PRD of a readability index for a certain corpus is 25, for example, the value of that readability index for the simplified versions of the texts is 25% lower than the value of the same index for the corresponding original versions (on average). The higher the mean value of the paired relative differences, the higher the level of complexity reduction achieved.

tive differences (PRD) of the corresponding readability index. The PRD were calculated according to Eq. 7.12, where $o_i(x)$ and $s_i(x)$ represent the value of the readability in-

dex *x* on the *i*th *original* text ($o_i(x)$), and the value of the readability index *x* on the *i*th simplified text ($s_i(x)$).

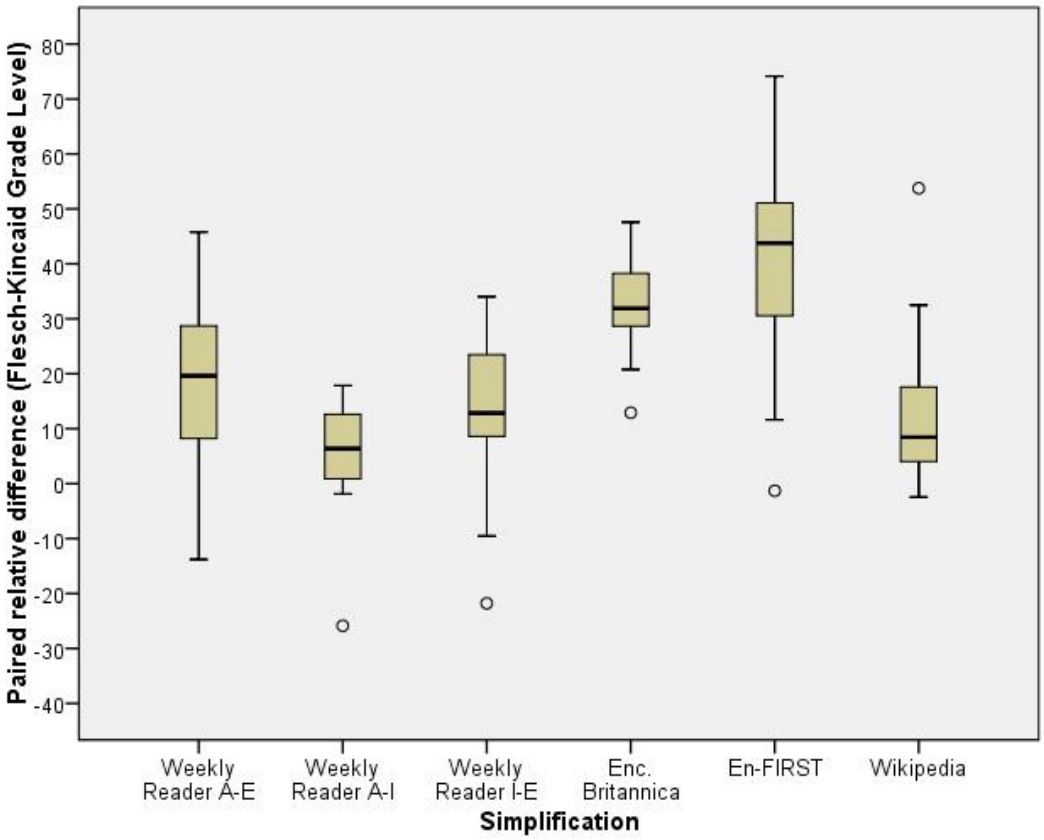
$$PRD = 100 - \frac{100 * s_i(x)}{o_i(x)} \tag{7.12}$$

It can be noted (Figure 7.1) that the level of simplification (measured by paired relative differences of SCI, LC, and SSR) achieved by automatic simplification (*Simplext (automatic)*) is much lower than the desired one achieved by manual simplification (*Simplext (manual)*). At the same time, the level of simplification achieved by the automatic simplification system built under the Simplext project (*Simplext (automatic)*) is very close to, and in terms of syntactic simplification measured by SCI even better than, the one achieved by manual simplification in the *FIRST* project (Figure 7.1). This indicates a possibility that some components of the automatic simplification system in Simplext (e.g. the syntactic simplification module) could be used for the syntactic simplification of texts for people with ASD. However, this possibility would need to be carefully investigated especially because the texts in *FIRST* and those in *Simplext (Original)* are not from the same domain and do not seem to have the same complexity (Table 7.13, Section 7.6.1).

The comparison of complexity reduction achieved by different simplification strategies for English are illustrated using the paired relative differences of the Kincaid-Flesch Grade Level (KFGL) for each pair of texts in each corpus (Figure 7.2).

The results (Figure 7.2) clearly indicate that simplification of texts aimed at people with ASD (*En-FIRST*) requires a higher level of complexity reduction than simplification of texts aimed at children (*Enc.Britannica*) or language learners (*WeeklyReader A-E*). At the same time, simplification of texts for children (*Enc.Britannica*) requires

Figure 7.2: Comparison of different text simplification strategies for English



The y-axis contains the paired relative differences (PRD) of the Flesch-Kincaid Grade Level (FKGL) index for each pair of texts in each of the six (sub-)corpora. The *WeeklyReader A-I* corresponds to a subcorpus of the WeeklyReader corpus, which contains only the texts aimed at advanced (*A*) and intermediate (*I*) language learners. Similarly, the *WeeklyReader A-E* and *Weekly Reader I-E* correspond to the subcorpora of the WeeklyReader corpus, which contain only texts aimed at advanced and elementary language learners (*A-E*), and only texts aimed at intermediate and elementary language learners (*I-E*). The height of the rectangle indicates the spread of the PRD on each (sub-)corpus, the horizontal line inside the rectangle indicates the mean, while the whiskers outside the rectangle indicate the smallest and largest observations which are not outliers. Outliers are presented with small circles beyond the whiskers. If the mean of the paired relative differences of the FKGL index for a certain corpus is 25, for example, the FKGL index of the simplified versions of the texts is for 25% lower than the same index on the corresponding original versions (on average). The higher the mean value of the paired relative differences, the higher level of complexity reduction achieved by the manual simplification.

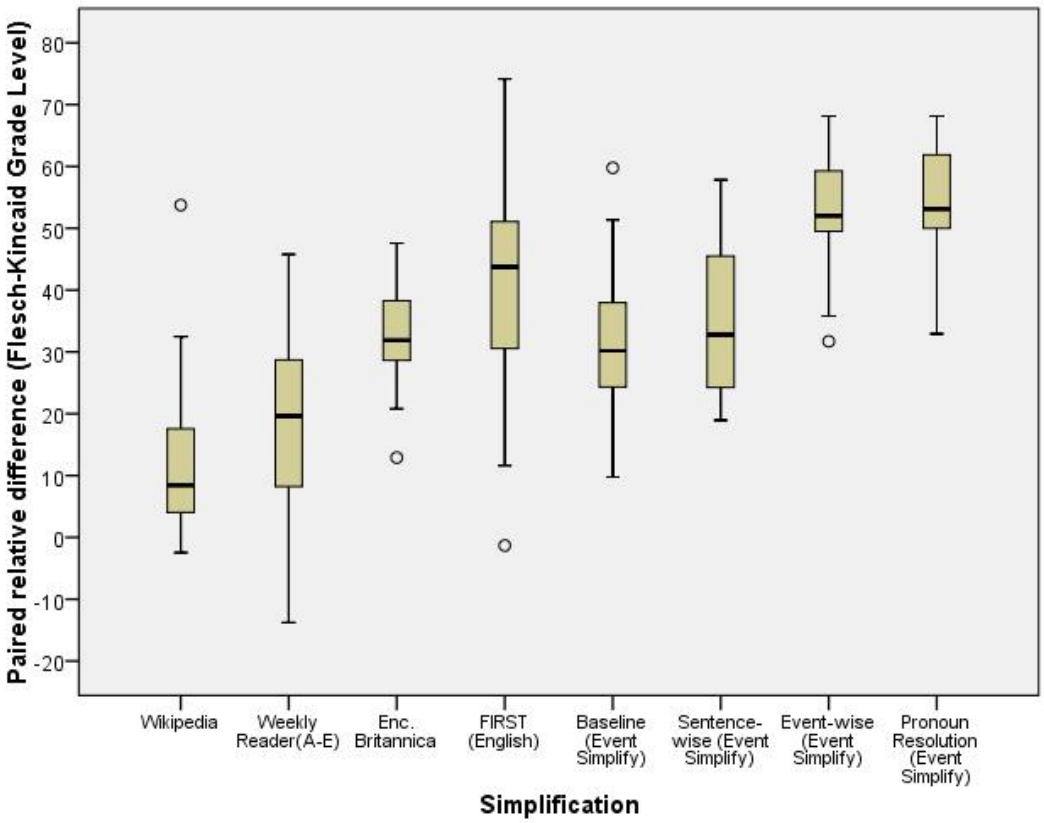
a higher level of complexity reduction than simplification of texts aimed at language learners of any level (*WeeklyReader A-E*, *WeeklyReader A-I*, *WeeklyReader I-E*).

Another interesting observation is that the differences between the complexity of the texts in English Wikipedia and their corresponding versions in Simple English Wikipedia (*Wikipedia*) seem to correspond only to the differences between the texts for language learners at the advanced and intermediate levels (*WeeklyReader A-I*). This raises the question as to whether the only existing large sentence-aligned corpus of original and simplified texts (based on the corresponding sentences of the English Wikipedia and the Simple English Wikipedia) represents good training material for building text simplification systems for any other target population than the language learners at the intermediate level.

7.6.3 Evaluation of our Text Simplification Systems

Readability indices can also be used for a quick (rough) comparison of different versions of the proposed ATS systems and their comparison with various manual simplification strategies (in terms of content reduction calculated using readability indices). Figure 7.3 illustrates this possibility on the example of the EventSimplify ATS system we proposed in Chapter 6. The baseline ATS system (which retains only the main clause of a sentence) and three versions of the EventSimplify system (sentence-wise simplification, event-wise simplification, and event-wise simplification with anaphoric pronoun resolution) are compared among themselves and with various previously mentioned manual simplification strategies. The comparison was done on the basis of complexity reduction calculated as a pair relative difference (eq. 7.12) of the Flesch-Kincaid Grade Level

Figure 7.3: EventSimplify vs. manual simplification in English



The y-axis contains the paired relative differences (PRD) of the Flesch-Kincaid Grade Level index for each pair of texts in each simplification version: B – the baseline system, S – the sentence-wise simplification by EventSimplify, E – the event-wise simplification by EventSimplify, and P – the event-wise simplification with anaphoric pronoun resolution by EventSimplify.

(FKGL) between the original text and the corresponding simplified version in each system.

The results presented in Figure 7.3 indicate that two of our EventSimplify simplification schemes (event-wise simplification and event-wise simplification with anaphoric pronoun resolution) achieve higher content reduction than any manual simplification strategy they were compared to (Wikipedia – aimed at broad audiences, Weekly Reader

– aimed at language learners, Enc. Britannica – aimed at children, and FIRST – aimed at people with ASD). The other two simplification schemes (baseline and sentence-wise simplification) seem to achieve content reduction similar to that in manual simplification of texts for children, lower than that in manual simplification of text for people with ASD, and higher than the content reduction achieved in manual simplification of texts for language learners (Weekly Reader) and wider audiences (Wikipedia).

Interestingly enough, the ranking of our four automatic simplification schemes by content reduction measured as paired relative differences of the Flesch-Kincaid Grade Level (FKGL) index seem to correspond perfectly to the ranking of those systems by human annotators in terms of sentence simplicity captured by the information relevance score (Table 7.15).

Table 7.15: Automatic vs. human evaluation of simplicity (content reduction)

Scheme	FKGL	Information Relevance
Baseline	-27.70% $\pm$ 12.51%	1.90 $\pm$ 0.64
Sentence-wise	-30.12% $\pm$ 13.93%	2.12 $\pm$ 0.61
Event-wise	-47.76% $\pm$ 13.91%	2.30 $\pm$ 0.54
Pronominal anaphora	-50.25% $\pm$ 12.59	2.39 $\pm$ 0.57

7.7 Summary

This chapter presented experiments into the suitability of using existing readability indices for automatic assessment of simplicity achieved by text simplification systems for English and Spanish. The first set of experiments supported the idea of using the se-

lected linguistically motivated features as a measure of text complexity (Section 7.3). The next set of experiments indicated a significant correlation between readability indices and the linguistically motivated features in both languages (Sections 7.4 and 7.5). Based on those findings, several possible uses of readability indices in text simplification were further highlighted in Section 7.6. Finally, our four simplification schemes proposed under the EventSimplify ATS system for English (Chapter 6) were compared among themselves and with various manual text simplifications strategies for English (Section 7.6.3).

CHAPTER 8

CONCLUSIONS

Text simplification (TS) is a relatively new research area which has a goal of transforming complex texts into their lexically and syntactically simpler variants. The benefits of text simplification are two-fold; it makes texts more accessible to wider audiences, and it improves the performance of various NLP systems. The focus of this thesis was on identifying and better understanding the main problems in automatic text simplification (ATS) and proposing new data-driven approaches to address them. The next three sections revisit the main research questions (Section 8.1), summarise the main findings of each chapter, comment on their potential impact on future text simplification studies (Section 8.2), and propose new research avenues (Section 8.3).

8.1 Research Questions Revisited

The extensive literature review presented in Chapters 2 and 3 identified four main problems in the current state-of-the-art ATS systems:

1. Parallel corpora for text simplification aimed at specific target populations are very scarce and limited in their size.

2. Automatic text simplification systems require either a large number of handcrafted simplification rules or large amounts of parallel data.

3. The existing ATS systems do not perform sufficient content reduction.

4. There is no well-established methodology for evaluating text simplification systems and comparing their performance.

With the aim of addressing those main problems in ATS systems, we formulated four research questions:

- **RQ 1:** Is it possible to adapt an already existing TS system aimed at a specific target audience to a TS system aimed at a different target population?

- **RQ 2:** Is it possible to build an ATS system which would not require large amounts of parallel data or handcrafted rules, but rather exploits some already existing NLP tools and can easily be adapted to different languages?

- **RQ 3:** Is it possible to build an ATS system which would, in addition to simplifying the given text, also perform significant content reduction by deleting irrelevant information?

- **RQ 4:** Could some of the already existing readability indices be used for the automatic evaluation of text simplification systems?

The first research question (**RQ 2**) was addressed in Chapter 4. The chapter focused on decision-making systems for sentence splitting and sentence deletion in text simplification systems for Spanish. The results indicated that the adaptation of an ATS system from one target audience to another is possible in the case of a sentence splitting decision-making module but not in the case of a sentence deletion decision-making module.

The second research question (**RQ 1**) was addressed in Chapters 5 and 6. The results of the experiments presented in Chapter 5 rejected the widespread assumption that the success of a PB-SMT approach largely depends on the size of the training and development datasets. They further showed how the sentence pairs in the training and development datasets can be filtered to improve the 'translation' performance, and achieve fair performance using the PB-SMT approach to TS even on small datasets. In Chapter 6, we proposed EventSimplify, a semantically motivated, event-based ATS system for news stories in English. The proposed system was built upon a state-of-the-art event extraction system. It does not require any parallel data, and it employs only a few handcrafted simplification rules which can easily be adapted to other languages. The adaptation of the system to other languages and domains mainly depends on the availability of a robust enough event extraction system for the required language and domain.

The third research question (**RQ 3**) was addressed in Chapters 4 and 6. The first part of Chapter 4, focused on building a decision-making system for sentence deletion in Spanish (Section 4.3), indicated that this is not a trivial task. Chapter 6 showed that our semantically motivated, event-based ATS system for news stories in English (EventSimplify) can successfully perform significant content reduction together with sentence simplification.

The fourth research question (**RQ 4**) was addressed in Chapter 7. We investigated whether some of the already existing readability formulae have a good correlation with the possible obstacles to reading comprehension and thus could be used for the automatic evaluation of simplicity achieved by text simplification systems. Our experiments

in English and Spanish indicated that there is a significant correlation between readability indices and the linguistically motivated features we proposed (Sections 7.4 and 7.5). Based on those findings, in Section 7.6, we suggested several possible uses of relative values of readability indices in the automatic evaluation of text simplification systems.

8.2 Original Contributions and their Impact

This thesis makes a number of novel contributions to text simplification by critically analysing the existing approaches, and proposing new ATS systems and new evaluation methods. In this section, we look back at the main contributions of each chapter and its envisioned impact on future text simplification studies.

In Chapter 4, we proposed a new feature set which leads to the state-of-the-art performance of two decision-making modules in ATS systems for Spanish: (1) classification of original sentences into those to be deleted and those to be kept during simplification; and (2) classification of original sentences into those to be split and those to be left unsplit during simplification. The main potential of these classification systems lies in enriching the state-of-the-art rule-based text simplification systems (such as the ATS system for Spanish proposed under the Simplext project, for example) if they are included at the beginning of the simplification pipeline. The proposed classification systems can eliminate unnecessary sentences (thus introducing a content reduction module which is currently not present in any of the rule-based systems) and detect the sentences which need to be split (and thus send them to a dedicated syntactic simplification module). The experiments into the adaptation of those two decision-making modules to different target populations and text genres opened a new research direction in text

simplification, which could help to overcome one of the main problems in current TS, the scarcity of parallel datasets aimed at specific target populations.

Chapter 5 presented several sets of experiments which led to a better understanding of the standard PB-SMT approach to text simplification. First, our results indicated that the type of the datasets (parallel or comparable) does not have any impact on the success of a standard PB-SMT model in text simplification. This finding is very encouraging given that one of the main problems of any data-driven approach to text simplification is the scarcity of parallel TS corpora which consist of original sentences and their corresponding manual simplifications. The compilation of comparable TS corpora should be an easier task than compilation of parallel TS corpora, as it requires less manual work and human expertise. Our next finding rejected the widespread assumption that the success of a PB-SMT approach largely depends on the size of the training and development datasets. The results indicated that the size of the datasets does not significantly influence the system's performance. This is particularly important as one of the main problems in TS is not only the scarcity of the parallel TS corpora but also the size of the existing data (usually about 1,000 sentence pairs or fewer). The results of the experiments conducted in Chapter 5 further indicated that the similarity between the original sentences and their corresponding manual simplifications in the training and development datasets has a strong impact on the performance of the system. This finding can be used to better model the standard PB-SMT systems for ATS by carefully selecting training and development datasets. Finally, we showed that BLEU is not a good measure of the performance of a standard PB-SMT model in ATS, as it mainly reflects the similarity between the original sentences and their simplified versions in the test set

and not the actual system's performance. While in cross-lingual MT a system which does not perform any translation/modification achieves a zero BLEU score, in monolingual MT a system which does not perform any translation/modification can achieve any BLEU score (high or low) depending on the test set used. In monolingual MT, the BLEU score of the system which does not perform any translation/modification on the input sentences is equal to the BLEU score between the original sentences and their reference/manual simplifications in the test set.

In Chapter 6, we proposed a new automatic text simplification system (EventSimplify) which simultaneously simplifies and reduces the content of a given text. The system does not require any parallel TS data nor large numbers of handcrafted simplification rules. It is semantically motivated and built upon a state-of-the-art event extraction system. The performance of the system is comparable to the state-of-the-art ATS systems in English (which require large parallel TS datasets), and it can be easily adapted to a different language under the condition that there is a robust enough event extraction system for that language.

Chapter 7 demonstrated that some of the already existing readability indices have a good correlation with the possible obstacles to reading comprehension and thus can be used for the automatic evaluation of simplicity achieved by text simplification systems. The experiments reported comparable results in English and Spanish. Based on those findings, we suggested several possible uses of readability indices in the automatic evaluation of text simplification systems. First, original and simplified texts can be compared in terms of readability indices in order to assess either the necessary complexity reduction (if comparing original texts with the manually simplified ones); or

the achieved complexity reduction (if comparing original texts with the automatically simplified ones). Second, the level of simplification achieved by different TS systems can be compared by using the relative differences of readability indices between original texts and their automatic simplifications performed by those TS systems. Third, automatically simplified texts can be compared with the manually simplified ones using readability indices, in order to assess whether the automatic simplification achieves the same level of simplification as the manual one. Finally, manually simplified texts can be compared with a 'gold standard' (easy-to-read texts which were originally written with the target population in mind) with the aim of assessing whether the manually simplified texts reach the simplicity of the 'gold standard' and thus comply with the easy-to-read standards. The envisaged use of the readability indices presented in Chapter 7 should provide a unified evaluation strategy for text simplification systems, which would enable a fairer comparison of their performance.

8.3 Future Work

The experiments presented in this thesis opened many avenues for further research. In this section, we will briefly present some of them.

The experiments presented in Chapter 5 indicated the possibility of better modelling the standard PB-SMT systems for ATS by carefully selecting training and development datasets, in order to improve the grammaticality and meaning preservation of the output sentences. We also showed that the size of the datasets does not significantly influence the system's performance. Those findings indicate that, by careful selection of sentence pairs for training and development datasets, the standard PB-SMT systems could per-

form fairly well in ATS in other domains and languages for which there is only a very limited amount of TS parallel data.

The results of the experiments presented in Chapter 5 also showed that the main limitations of PB-SMT systems for ATS lie in the lack of sentence splitting and content reduction, leading to low human scores for simplicity of the output. At the same time, the EventSimplify ATS system presented in Chapter 6 achieved high scores for grammaticality, meaning preservation and simplicity of its output, but it does not perform any lexical simplification. In future, we could overcome the main limitations of these two approaches (the PB-SMT approach and the event-based approach) by combining the best PB-SMT systems for TS with the proposed EventSimplify system. This combination would result in an ATS system which performs lexical and syntactic simplification with significant content reduction.

The EventSimplify system (Chapter 6), on its own, could be improved by slightly modifying the simplification algorithms in order to avoid recurring errors identified during the error analysis (e.g. loss of timeline). The evaluation of the system could be enriched by human assessment of the whole texts which would point out possible problems in text coherence after content reduction. We could also perform an evaluation of the simplified versions of the texts on which the event-extraction system was trained, i.e. texts with 'gold standard' events. The human scores for grammaticality, meaning preservation, and simplicity obtained on those texts would represent an upper bound of our simplification system.

Finally, the collected datasets with human scores for grammaticality, meaning preservation and simplicity of simplified sentences could be used for training decision-making

systems which would classify automatically simplified sentences into three categories: (1) correct sentences ready to be presented to the users; (2) sentences which require minimal post-editing (automatic or manual) in order to be corrected and presented to the users; and (3) incorrect sentences (e.g. sentences whose original meanings were completely changed) which need to be discarded. Our initial experiments following this idea can be found in one of our previously published studies (Štajner et al., 2014b).

BIBLIOGRAPHY

Aduriz, I., Agirre, E., Aldezabal, I., Alegria, I., Ansa, O., Arregi, X., Arriola, J., Artola, X., de Ilarraza, A. D., Ezeiza, N., Gojenola, K., Maritxalar, M., Oronoz, M., Sarasola, K., Soroa, A., Urizar, R., and Urkia, M. (1998). A framework for the automatic processing of Basque. In *Proceedings of the LREC Workshop on Lexical Resources for Minority Languages*.

Aduriz, I., I., A., Alegria, I., Arriola, J., Artola, X., de Ilarraza, A. D., Ezeiza, N., and Gojenola, K. (2003). Finite State Applications for Basque. In *Proceedings of the EACL Workshop on Finite-State Methods in Natural Language Processing*, pages 3–11.

Alegria, I., Ezeiza, N., Fernandez, I., and Urizar, R. (2003). Named entity recognition and classification for texts in Basque. In *II Jornadas de Tratamiento y Recuperación de Información, JOTRI, Madrid*.

Allen, D. (2009). A study of the role of relative clauses in the simplification of news texts for learners of English. *System*, 37(4):585–599.

Aluísio, S., Specia, L., Gasperin, C., and Scarton, C. (2010). Readability assessment for text simplification. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications (IUNLPBEA)*, pages 1–9, Stroudsburg, PA, USA. Association for Computational Linguistics.

Aluísio, S. M. and Gasperin, C. (2010). Fostering Digital Inclusion and Accessibility: The PorSimples Project for Simplification of Portuguese Texts. In *Proceedings of the NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas (YIWCALA)*, pages 46–53. ACL.

Aluísio, S. M., Specia, L., Pardo, T. A. S., Maziero, E. G., Caseli, H. M., and Fortes, R. P. M. (2008). A corpus analysis of simple account texts and the proposal of simplification strategies: first steps towards text simplification systems. In *Proceedings of the 26th annual ACM international conference on Design of communication (SIGDOC)*, pages 15–22. ACM.

Angrosh, M. and Siddharthan, A. (2014). Text simplification using synchronous dependency grammars: Generalising automatically harvested rules. In *Proceedings of the 8th International Natural Language Generation Conference (INGL)*, pages 16–25.

Anula, A. (2007). Tipos de textos, complejidad lingüística y facilicitación lectora. In *Actas del Sexto Congreso de Hispanistas de Asia*, pages 45–61.

Aranzabe, M. J., Díaz De Ilarraza, A., and González, I. (2012). First Approach to Automatic Text Simplification in Basque. In *Proceedings of the first Natural Language Processing for Improving Textual Accessibility Workshop (NLP4ITA)*, pages 1–8.

Balota, D., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., and Yap, M. J. (2004). Visual word recognition of single-syllabe words. *Journal of Experimental Psychology: General*, 133(2):283–316.

Barlacchi, G. and Tonelli, S. (2013). ERNESTA: A sentence simplification tool for childrens stories in Italian. In *Computational Linguistics and Intelligent Text Processing*, pages 476–487. Springer.

Barzilay, R. and Elhadad, N. (2003). Sentence alignment for monolingual comparable corpora. In *Proceedings of the Conference on Empirical methods in Natural Language Processing (EMNLP)*, pages 25–32. ACL.

Bautista, S., Gervás, P., and Madrid, I. (2009). Feasibility Analysis for SemiAutomatic Conversion of Text to Improve Readability. In *Proceedings of the 2nd International Conference on Information and Communication Technology and Accessibility (ICTA)*, pages 33–40.

Bautista, S., León, C., Hervás, R., and Gervás, P. (2011). Empirical Identification of Text Simplification Strategies for Reading-Impaired People. In *European Conference for the Advancement of Assistive Technology (AAATE)*, pages 567–574.

Beigman Klebanov, B., Knight, K., and Marcu, D. (2004). Text simplification for information-seeking applications. In *On the Move to Meaningful Internet Systems*, volume 3290 of *LNCS*, pages 735–747. Springer Berlin Heidelberg.

Bick, E. (2000). *The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. PhD thesis, Aarhus University.

Biderman, M. (2005). *Dicionário Iustrado de Português*. Editora Ática, São Paulo.

Biran, O., Brody, S., and Elhadad, N. (2011). Putting it Simply: a Context-Aware Approach to Lexical Simplification. In *Proceedings of the 49th Annual Meeting of the*

Association for Computational Linguistics: Human Language Technologies (ACL-HLT), pages 496–501. ACL.

Bohnet, B. (2009). Efficient parsing of syntactic and semantic dependency structures. In *Proceedings of the Conference on Natural Language Learning (CoNLL)*, pages 67–72. Association for Computational Linguistics.

Bohnet, B., Langjahr, A., and Wanner, L. (2000). A development environment for MTT-based sentence generators. *Revista de la Sociedad Española para el Procesamiento del Lenguaje Natural*, 26:35–36.

Bormuth, J. R. (1966). Readability: A new approach. *Reading research quarterly*, 1:79–132.

Bosque Muñoz, I. and Demonte Barreto, V. (1999). *Gramática Descriptiva de la Lengua Española*. Real Academia Española.

Bott, S., Rello, L., Drndarevic, B., and Saggion, H. (2012a). Can Spanish Be Simpler? LexSiS: Lexical Simplification for Spanish. In *Proceedings of COLING*, pages 357–374, Mumbai, India.

Bott, S. and Saggion, H. (2011). Spanish Text Simplification: An Exploratory Study. *Revista de la Sociedad Española para el Procesamiento del Lenguaje Natural*, 47:87–95.

Bott, S., Saggion, H., and Mille, S. (2012b). Text Simplification Tools for Spanish. In *Proceedings of LREC*, pages 1665–1671.

Briscoe, E. and Carroll, J. (1993). Generalized probabilistic LR parsing of natural language (corpora) with unification-based grammars. *Computational Linguistics*, 19(1):25–60.

Briscoe, E., Carroll, J., and Watson, R. (2006). The second release of the RASP system. In *Proceedings of the COLING/ACL Interactive Presentation Session*, volume 6, pages 77–80. ACL.

Brouwer, R. H. M. (1963). Onderzoek naar de leesmoeilijkheden van nederlands proza. *Pedagogische studiën*, 40:454–464.

Brouwers, L., Bernhard, D., Ligozat, A., and François, T. (2014). Syntactic sentence simplification for french. In *Proceedings of the EACL Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR), Gothenburg, Sweden*, pages 47–56.

Burstein, J., Shore, J., Sabatini, J., Lee, Y., and Ventura, M. (2007). The automated text adaptation tool. In *Proceedings of NAACL HLT Demonstration Program*, pages 3–4. ACL.

Candito, M., Crabbé, B., and Denis, P. (2010). Statistical French dependency parsing: treebank conversion and first results. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC)*, pages 1840–1847.

Canning, Y., Tait, J., Archibald, J., and Crawley, R. (2000). Cohesive Generation of Syntactically Simplified Newspaper Text. In *Proceedings of the Third International Workshop on Text, Speech and Dialogue (TSD)*, pages 145–150, London, UK.

Carroll, J., Minnen, G., Canning, Y., Devlin, S., and Tait, J. (1998). Practical Simplification of English Newspaper Text to Assist Aphasic Readers. In *Proceedings of AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology*, pages 7–10.

Carroll, J., Minnen, G., Pearce, D., Canning, Y., Devlin, S., and Tait, J. (1999). Simplifying text for language-impaired readers. In *Proceedings of the 9th Conference of the European Chapter of the ACL (EACL)*, pages 269–270.

Caseli, H. M., Pereira, T. F., Specia, L., Pardo, T. A. S., Gasperin, C., and Aluísio, S. M. (2009). Building a Brazilian Portuguese parallel corpus of original and simplified texts. In *Advances in Computational Linguistics. Research in Computer Science*, volume 41, pages 59–70.

Chandrasekar, R. (1994). *A Hybrid Approach to Machine Translation using Man Machine Communication*. PhD thesis, Tata Institute of Fundamental Research/University of Bombay, Bombay.

Chandrasekar, R., Doran, C., and Srinivas, B. (1996). Motivations and methods for text simplification. In *In Proceedings of the 16th International Conference on Computational Linguistics (COLING)*, pages 1041–1044.

Chandrasekar, R. and Srinivas, B. (1996). Automatic Induction of Rules for Text Simplification. Technical report, Institute for Research in Cognitive Science, University of Pennsylvania.

Chandrasekar, R. and Srinivas, B. (1997). Automatic induction of rules for text simplification. *Knowledge-Based Systems*, 10(3):183 – 190.

Chomsky, N. (1986). *Knowledge of language: its nature, origin, and use.* Greenwood Publishing Group, Santa Barbara, California.

Clarke, J. and Lapata, M. (2006). Models for sentence compression: A comparison across domains, training requirements and evaluation measures. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics (ACL)*, pages 377–384. ACL.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70:213–220.

Cohen, W. (1995). Fast Effective Rule Induction. In *Proceedings of the Twelfth International Conference on Machine Learning (ICML)*, pages 115–123.

Cohn, T.and Lapata, M. (2009). Sentence compression as tree transduction. *Journal of Artificial Intelligence Research (JAIR)*, 34:637–674.

Coleman, E. B. (1971). *Developing a technology of written instruction: some determiners of the complexity of prose*. Teachers College Press, Columbia University, New York.

Coleman, M. and Liau, T. L. (1975). A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283–284.

Coster, W. and Kauchak, D. (2011a). Learning to Simplify Sentences Using Wikipedia. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1–9. ACL.

Coster, W. and Kauchak, D. (2011b). Simple English Wikipedia: a new text simplification task. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 665–669. ACL.

Cuetos, F., Domínguez, A., and de Vega, M. (1997). El efecto de la polisemia: ahora lo ves otra vez. *Cognitiva*, 9(2):175–194.

Cunningham, H., Wilks, Y., and Gaizauskas, R. (1996). GATE – A general architecture for text engineering. In *Proceedings of the 16th Conference on Computational Linguistics (COLING)*, pages 29–30.

Curran, J. R., Clark, S., and Bos, J. (2007). Linguistically motivated large-scale NLP with C&C and Boxer. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL) on Interactive Poster and Demonstration Sessions*, pages 33–36. ACL.

Dale, E. and Chall., J. S. (1948). A Formula for Predicting Readability. *Educational Research Bulletin*, 27(1):11–20.

De Belder, J. and Moens, M. (2010). Text simplification for children. In *Proceedings of the SIGIR workshop on accessible search systems*, pages 19–26.

de Marneffe, M., MacCartney, B., and Manning, C. (2006). Generating typed depen-

dency parses from phrase structure parses. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pages 449–454.

DellOrletta, F., Montemagni, S., and Venturi, G. (2011). Read-it: Assessing readability of Italian texts with a view to text simplification. In *Proceedings of the 2nd Workshop on Speech and Language Processing for Assistive Technologies (SLPAT)*, pages 73–83.

Denis, P. and Sagot, B. (2009). Coupling an annotated corpus and a morphosyntactic lexicon for state-of-the-art POS tagging with less human effort. In *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation (PACLIC), Volume 1*, pages 110–119.

Denkowski, M. and Lavie, A. (2011). Meteor 1.3: Automatic Metric for Reliable Optimization and Evaluation of Machine Translation Systems. In *Proceedings of the EMNLP Workshop on Statistical Machine Translation*, pages 85–91.

Deschacht, K. and Moens, M.-F. (2009). The Latent Words Language Model. In *Proceedings of the 18th Annual Belgian-Dutch Conference on Machine Learning*.

Devlin, S. (1999). *Simplifying natural language text for aphasic readers*. PhD thesis, University of Sunderland, UK.

Devlin, S. and Tait, J. (1998). The use of a psycholinguistic database in the simplification of text for aphasic readers. *Linguistic Databases*, pages 161–173.

Devlin, S. and Unthank, G. (2006). Helping aphasic people process online information.

In *Proceedings of the 8th international ACM SIGACCESS conference on Computers and accessibility (Assers)*, pages 225–226. ACM.

Ding, Y. and Palmer, M. (2005). Machine translation using probabilistic synchronous dependency insertion grammars. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL)*, pages 541–548. ACL.

Doddington, G. (2002). Automatic evaluation of machine translation quality using n-gram cooccurrence statistics. In *Proceedings of the 2nd Conference on Human Lnaguage Technology Research*, pages 138–145, San Diego, USA.

Dornescu, I., Evans, R., and Orasan, C. (2013). A Tagging Approach to Identify Complex Constituents for Text Simplification. In *Proceedings of Recent Advances in Natural Language Processing (RANLP)*, pages 221 – 229, Hissar, Bulgaria.

Douma, W. (1960). De leesbaarheid van landbouwbladen: een onderzoek naar en een toepassing van leesbaarheidsformules. *Bulletin*, 17.

Drndarević, B. and Saggion, H. (2012). Reducing Text Complexity through Automatic Lexical Simplification: an Empirical Study for Spanish. *Revista de la Sociedad Española para el Procesamiento del Lenguaje Natural*, 49:13–20.

Drndarević, B., Štajner, S., Bott, S., Bautista, S., and Saggion, H. (2013). Automatic Text Simplication in Spanish: A Comparative Evaluation of Complementing Components. In *Proceedings of the 12th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing)*, pages 488–500.

Drndarevic, B., Štajner, S., and Saggion, H. (2012). Reporting Simply: A Lexical Simplification Strategy for Enhancing Text Accessibility. In *Proceedings of the Easy-to-read on the Web Symposium*.

DuBay, W. H. (2004). The Principles of Readability. *Impact Information*.

Ehrlich, M., Remond, M., and Tardieu, H. (1999). Processing of anaphoric devices in young skilled and less skilled comprehenders: Differences in metacognitive monitoring. *Reading and Writing*, 11(1):29–63.

Elworthy, D. (1994). Does baum-welch re-estimation help taggers? In *Proceedings of the 4th ACL conference on Applied Natural Language Processing*, pages 53–58.

Evans, R. J. (2011). Comparing methods for the syntactic simplification of sentences in information extraction. *Literary and Linguistic Computing*, 26(4):371–388.

Ezeiza, N. (2002). *CORPUSAK USTIATZEKO TRESNA LINGUISTIKOAK. Euskararen etiketatzaile morfosintaktiko sendo eta malgua.* PhD thesis, University of the Basque Country.

Fajardo, I., Ávila, V., Ferrer, A., Tavares, G., Gómez, M., and Hernández, A. (2014). Easy-to-read texts for students with intellectual disability: linguistic factors affecting comprehension. *Journal of Applied Research in Intellectual Disabilities*, 27(3):212–225.

Feblowitz, D. and Kauchak, D. (2013). Sentence simplification as tree transduction. In *Proceedings of the 2n Workshop on Predicting and Improving Text Readability for Targe Reader Populations (PITR)*, pages 1–10.

Fellbaum, C. (2010). Wordnet. In Poli, R., Healy, M., and Kameas, A., editors, *Theory and Applications of Ontology: Computer Applications*, pages 231–243. Springer Netherlands.

Feng, L. (2009). Automatic readability assessment for people with intellectual disabilities. In *ACM SIGACCESS Accessibility and Computing*, number 93, pages 84–91. ACM.

Feng, L., Elhadad, N., and Huenerfauth, M. (2009). Cognitively motivated features for readability assessment. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 229–237. ACL.

Flesch, R. (1949). *The art of readable writing*. Harper, New York.

Francois, T. and Watrin, P. (2011). On the Contribution of MWE-based Features to a Readability Formula for French as a Foreign Language. In *Proceedings of Recent Advances in Natural Language Processing (RANLP)*, pages 441–447.

Freyhoff, G., Hess, G., Kerr, L., Tronbacke, B., and Van Der Veken, K. (1998). *Make it Simple, European Guidelines for the Production of Easy-toRead Information for People with Learning Disability*. ILSMH European Association, Brussels.

Gaizauskas, R., Foster, J., Wilks, Y., Arundel, J., Clough, P., and Piao, S. (2001). The METER Corpus: A corpus for analysing journalistic text reuse. In *Proceedings of Corpus Linguistics Conference*, pages 214–223. Lancaster University Centre for Computer Corpus Research on Language.

Gasperin, C., Specia, L., Pereira, T., and Aluísio, S. (2009). Learning When to Simplify Sentences for Natural Text Simplification. In *Proceedings of the Encontro Nacional de Inteligncia Artificial (ENIA), Bento Gonalves, Brazil*, pages 809–818.

Geman, S. and Johnson, M. (2002). Dynamic programming for parsing and estimation of stochastic unification-based grammars. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia*, pages 279–286.

Gernsbacher, M. A. and Faust, M. (1991). The mechanism of suppression: A component of general comprehension skill. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 17:245–262.

Glanzer, M. and Bowles, N. (1976). Analysis of the word frequency effect in recognition memory. *Journal of Experimental Psychology: Human Learning and Memory*, 2:21–31.

Glavaš, G. and Šnajder, J. (2013). Exploring coreference uncertainty of generically extracted event mentions. In *Proceedings of 14th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing)*, pages 408–422. Springer.

Glavaš, G. and Šnajder, J. (2014). Construction and evaluation of event graphs. *Natural Language Engineering*, FirstView:1–46.

Glavaš, G. and Štajner, S. (2013). Event-Centered Simplication of News Stories. In *Pro-*

ceedings of the Student Workshop held in conjunction with RANLP, Hissar, Bulgaria, pages 71–78.

Gómez, P. (2011). Identificación de dificultades de comprensión lectora en el uso de internet en jóvenes con discapacidad intelectual para el desarrollo de un periódico digital. In *Poster presented at the XV Congreso Nacional y I Internacional de Modelos de Investigación Educativa, Madrid*.

Gonzalez-Dios, I., Aranzabe, M., Díaz de Ilarraza, A., and Salaberri, H. (2014). Simple or Complex? Assessing the readability of Basque Texts. In *Proceedings of the COLING, the 25th International Conference on Computational Linguistics: Technical Papers, Dublin, Ireland*, pages 334–344.

Grover, C., Matheson, C., Mikheev, A., and Moens, M. (2000). LT TTT – A Flexible Tokenisation Tool. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC), Athens, Greece*.

Gunning, R. (1952). *The technique of clear writing*. McGraw-Hill, New York.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18.

Hall, M. A. (1999). *Correlation-based Feature Selection for Machine Learning*. PhD thesis, The University of Waikato. Hamilton, New Zealand.

Hall, M. A. and Smith, L. A. (1998). Practical feature subset selection for machine

learning. In *Proceedings of the 21st Australasian Computer Science Conference (ACSC)*, pages 181–191. Berlin: Springer.

Ian H. Witten, E. F. (2005). *Data mining: practical machine learning tools and techniques*. Morgan Kaufmann Publishers.

Inui, K., Fujita, A., Takahashi, T., Iida, R., and Iwakura, T. (2003). Text simplification for reading assistance: a project note. In *Proceedings of the second international workshop on Paraphrasing (PARAPHRASE)*, volume 16, pages 9–16. ACL.

Janczura, G., Castilho, G., and Rocha, N. (2007). Normas de concretude para 909 palavras da língua portuguesa. *Psicologia: Teoria e Pesquisa*, 23(2):195–204.

Jastrzembski, J. (1981). Multiple meaning, number or related meanings, frequency of occurrence and the lexicon. *Cognitive Psychology*, 13:278–305.

Joshi, A. (1985). Tree adjoining grammars: how much context sensitivity is required to provide a reasonable structural description. In Dowty, D., Karttunen, I., and Zwicky, A., editors, *Natural Language Parsing*. Cambridge University Press.

Joshi, A. K. and Srinivas, B. (1994). Disambiguation of super parts of speech (or supertags): Almost parsing. In *Proceedings of the 15th Conference on Computational Linguistics (COLING)*, volume 1, pages 154–160. ACL.

Jurafsky, D. and Martin, J. H. (2008). *Speech and Language Processing*. Prentice Hall, 2 edition.

Kamp, H. (1981). A theory of truth and semantic representation. In Groenendijk, J., Janssen, T., Stokhof, B., and Stokho, M., editors, *Formal methods in the study of language, Part 1*, volume 136 of *Mathematical Centre tracts*. Mathematisch Centrum Amsterdam.

Karlsson, F., Voutilainen, A., Heikkila, J., and Anttila, A. (1995). *Constraint Grammar, A Language-independent System for Parsing Unrestricted Text*. Mouton de Gruyter.

Karreman, J., van der Geest, T., and Buursink, E. (2007). Accessible website content guidelines for users with intellectual disabilities. *Journal of Applied Research in Intellectual Disabilities*, 20:510–518.

Kauchak, D. (2013). Improving text simplification language modeling using unsimplified text data. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1537–1546, Sofia, Bulgaria. ACL.

Keerthi, S. S., Shevade, S. K., Bhattacharyya, C., and Murthy, K. R. K. (2001). Improvements to Platt's SMO Algorithm for SVM Classifier Design. *Neural Computation*, 13(3):637–649.

Kern, R. P. (2004). *Usefulness of readability formulas for achieving Army readability objectives: Research and state-of-the-art applied to the Army's problem (NTIS No. AD A086 408/2)*. U.S. Army Reserach Institute, Fort Benjamin Harrison.

Kincaid, J. P., Fishburne, R. P., Rogers, R. L., and Chissom, B. S. (1975). Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch

Reading Ease Formula) for Navy enlisted personnel. Research branch report 8-75, Naval Air Station Memphis, Millington, TN.

Klein, D. and Manning, C. (2003a). Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics (ACL)*, volume 1, pages 423–430. ACL.

Klein, D. and Manning, C. D. (2003b). Fast exact inference with a factored model for natural language parsing. In *Advances in Neural Information Processing Systems*, volume 15, pages 3–10.

Knight, K. and Marcu, D. (2002). Summarization beyond sentence extraction: A probabilistic approach to sentence compression. *Artificial Intelligence*, 139:91–107.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 177–180. ACL.

Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical phrase-based translation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (NAACL)*, volume 1, pages 48–54. ACL.

Kučera, H. and Francis, W. (1967). *Computational analysis of present-day American English*. University Press of New England.

Landis, J. and Koch, G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159–174.

Lavelli, A., Hall, J., Nilsson, J., and Nivre, J. (2009). MaltParser at the EVALITA 2009 Dependency Parsing Task. In *Proceedings of EVALITA*.

Lavie, A. and Denkowski, M. (2009). The METEOR Metric for Automatic Evaluation of Machine Translation. *Machine Translation*, 23:105–115.

Lee, H., Peirsman, Y., Chang, A., Chambers, N., Surdeanu, M., and Jurafsky, D. (2011). Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 28–34. ACL.

Li, Y., Zaragoza, H., Herbrich, R., ShaweTaylor, J., and Kandola, J. (2002). The perceptron algorithm with uneven margins. In *Proceedings of the 9th International Conference on Machine Learning (ICML)*, pages 379–386.

Lin, D. (1998a). Automatic retrieval and clustering of similar words. In *Proceedings of the 17th international conference on Computational linguistics (COLING)*, volume 2, pages 768–774. ACL.

Lin, D. (1998b). Dependency-Based Evaluation of MINIPAR. In Abeillé, A., editor, *Treebanks. Building and Using Parsed Corpora (Part II)*, Text, Speech and Language Technology, pages 317–329. Springer Netherlands.

Martos, J., Freire, S., González, A., Gil, D., and Sebastian, M. (2012). D2.1: Functional

requirements specifications and user preference survey. Technical report, FIRST project.

McLaughlin, G. H. (1969). SMOG grading: A new readability formula. *Journal of Reading*, 12(8):639–646.

Mencap (2002). Am I making myself clear? Mencaps guidelines for accessible writing.

Miller, G., Beckwith, R., Fellbaum, C., Gross, D., Miller, K., and Tengi, R. (1993). *Five Papers on Word-Net*. Princeton University, Princeton, N.J.

Mitkov, R. and Štajner, S. (2014). The Fewer, the Better? A Contrastive Study about Ways to Simplify. In *Proceedings of the COLING workshop on Automatic Text Simplification – Methods and Applications in the Multilingual Society (ATS-MA), Dublin, Ireland*, pages 30–40.

Morgan, M. F. and Moni, K. B. (2008). Meeting the challenge of limited literacy resources for adolescents and adults with intellectual disabilities. *British Journal of Special Education*, 35(2):92–101.

Napoles, C. and Dredze, M. (2010). Learning simple Wikipedia: a cogitation in ascertaining abecedarian language. In *Proceedings of the NAACL HLT Workshop on Computational Linguistics and Writing: Writing Processes and Authoring Aids (CL&W)*, pages 42–50. ACL.

Narayan, S. and Gardent, C. (2014). Hybrid simplification using deep semantics and machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 435–445.

Nomura, M., Skat Nielsen, G., and Tronbacke, B. (1997). Guidelines for easy-to-read materials. Technical report, IFLA, Library Services to People with Special Needs Section.

Norbury, C. (2005). Barking up the wrong tree? Lexical ambiguity resolution in children with language impairments and autistic spectrum disorders. *Journal of Experimental Child Psychology*, 90:142–171.

Och, F. (2003). Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 160–167.

Och, F. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Orasan, C., Evans, R., and Dornescu, I. (2013). Text simplification for people with autistic spectrum disorders. In D. Tufis, V. R. and Forascu, C., editors, *Towards Multilingual Europe 2020: A Romanian Perspective*, pages 287–312. Romanian Academy Publishing House, Bucharest.

Pan, Z. and Kosicki, G. M. (1993). Framing analysis: An approach to news discourse. *Political communication*, 10(1):55–75.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318.

Pardo, T. and Nunes, M. (2006). Review and Evaluation of DiZer – An Automatic Discourse Analyzer for Brazilian Portuguese. In *Computational Processing of the*

Portuguese Language (PROPOR), LNCS 3960, pages 180–189. Springer Berlin Heidelberg.

Petersen, S. E. and Ostendorf, M. (2007). Text Simplification for Language Learners: A Corpus Analysis. In *Proceedings of Workshop on Speech and Language Technology for Education (SLaTE)*.

Petersen, S. E. and Ostendorf, M. (2009). A machine learning approach to reading level assessment. *Computer Speech & Language*, 23(1):89 – 106.

Pianta, E., Girardi, C., and Zanoli, R. (2008). The TextPro tool suite. In *Proceedings of the 6th Lnaguage Resources and Evaluation Conference (LREC)*.

PlainLanguage (2011). Federal plain language guidelines.

Platt, J. C. (1998). Fast Training of Support Vector Machines using Sequential Minimal Optimization. In Schoelkopf, B., B. C. and Smola, A., editors, *Advances in Kernel Methods Support Vector Learning*, pages 185–210. MIT Press.

Pustejovsky, J., Castano, J., Ingria, R., Sauri, R., Gaizauskas, R., Setzer, A., Katz, G., and Radev, D. (2003). TimeML: Robust Specification of Event and Temporal Expressions in Text. *New Directions in Question Answering*, 3:28–34.

Quinlan, P. (1992). *The Oxford Psycholinguistic Database*. Oxford University Press.

Quinlan, R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Mateo, CA.

Quirk, R., Greenbaum, S., Leech, G., and Svartvik, J. (1985). *A Comprehensive Grammar of the English Language*. Longman Inc. New York.

Rello, L. (2012). DysWebxia: A Model to Improve Accessibility of the Textual Web for Dyslexic Users. In *ACM SIGACCESS Accessibility and Computing.*, number 102, pages 41–44. ACM, New York, NY, USA.

Rello, L., Baeza-Yates, R., Dempere, L., and Saggion, H. (2013). Frequent words improve readability and short words improve understandability for people with dyslexia. In *Human-Computer Interaction INTERACT 2013 (Part 4)*, LNCS, pages 203–219. Springer Berlin Heidelberg.

Roll, M., Frid, J., and Horne, M. (2007). Measuring syntactic complexity in spontaneous spoken Swedish. *Language and Speech*, 50(2).

Rosen, S. (1999). The syntactic representation of linguistic events. *Glot International*, 4(2):3–11.

Ruiter, M. B., Rietveld, T. C. M., Cucchiarini, C., Krahmer, E. J., and Strik, H. (2010). Human Language Technology and communicative disabilities: Requirements and possibilities for the future. In *Proceedings of the the seventh international conference on Language Resources and Evaluation (LREC)*, pages 2839–2846.

Rybing, J., Smith, C., and Silvervarg, A. (2010). Towards a Rule Based System for Automatic Simplification of Texts. In *The Third Swedish Language Technology Conference (SLTC)*.

Saggion, H., Gómez Martínez, E., Etayo, E., Anula, A., and Bourg, L. (2011). Text Simplification in Simplext: Making Text More Accessible. *Revista de la Sociedad Española para el Procesamiento del Lenguaje Natural*, 47:341–342.

Salton, G., Wong, A., and Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.

Schabes, Y., Abeillé, A., and Joshi, A. (1988). Parsing Strategies With 'Lexicalized' Grammars: Application To Tree Adjoining Grammars. In *Proceedings of the 12th International Conference on Computational Linguistics (COLING), Budapest, Hungary*, volume 2, pages 578–583.

Schwarm, S. E. and Ostendorf, M. (2005). Reading Level Assessment Using Support Vector Machines and Statistical Language Models. In *Proceedings of the 43rd annual meeting of the Association of Computational Linguistics (ACL)*, pages 523–530.

Shapiro, A. and Milkes, A. (2004). Skilled readers make better use of anaphora: a study of the repeated-name penalty on text comprehension. *Electronic Journal of Research in Educational Psychology*, 2(2):161–180.

Shardlow, M. (2014). Out in the Open: Finding and Categorising Errors in the Lexical Simplification Pipeline. In *Proceedings of Language Resources and Evaluation Conference (LREC)*, pages 1583–1590.

Siddharthan, A. (2002). An Architeture for a Text Simplification System. In *Proceedings of the Language Engineering Conference (LEC)*, pages 64–71. IEEE Computer Society Washington, DC, USA.

Siddharthan, A. (2006). Syntactic simplification and text cohesion. *Research on Language & Computation*, 4(1):77–109.

Siddharthan, A. (2010). Complex Lexico-Syntactic Reformulation of Sentences using Typed Dependency Representations. In *Proceedings of the 6th International Natural Language Generation Conference (INGL)*, pages 125–133.

Siddharthan, A. (2011). Text simplification using typed dependencies: a comparison of the robustness of different generation strategies. In *Proceedings of the 13th European Workshop on Natural Language Generation (ENLG)*, pages 2–11.

Siddharthan, A. and Angrosh, M. (2014). Hybrid text simplification using synchronous dependency grammars with hand-written and automatically harvested rules. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL), Gothenburg, Sweden*, pages 722–731.

Smith, D. A. and Eisner, J. (2006). Quasi-synchronous grammars: Alignment by soft projection of syntactic dependencies. In *Proceedings of the HLT-NAACL Workshop on Statistical Machine Translation*, pages 23–30. ACL.

Smith, E. A. and Senter, R. J. (1967). Automated Readability Index. Technical report, Aerospace Medical Research Laboratories, Wright-Patterson Air Force Base, Ohio.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A Study of Translation Edit Rate with Targeted Human Annotation,". In *Proceedings of Association for Machine Translation in the Americas*.

Snover, M., Madnani, N., Dorr, B., and Schwartz, R. (2009). Fluency, adequacy, or HTER? Exploring different human judgments with a tunable MT metric. In *Proceedings of the Fourth Workshop on Statistical Machine Translation, Athens, Greece*, pages 259–268.

Spaulding, S. (1956). A Spanish Readability Formula. *Modern Language Journal*, 40:433–441.

Specia, L. (2010). Translating from complex to simplified sentences. In *Proceedings of the 9th international conference on Computational Processing of the Portuguese Language (PROPOR)*, pages 30–39. Springer-Verlag Berlin, Heidelberg.

Stolcke, A. (2002). SRILM - an Extensible Language Modeling Toolkit. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, pages 901–904.

Van Dijk, T. A. (1985). Structures of news in the press. In *Discourse and Communication*, pages 69–93. Berlin: de Gruyter.

van Oosten, P., Tanghe, D., and Hoste, V. (2010). Towards an Improved Methodology for Automated Readability Prediction. In *Proceedings of the seventh international conference on language resources and evaluation (LREC),Valletta, Malta*, pages 775–782. European Language Resources Association (ELRA).

Vickrey, D. and Koller, D. (2008). Sentence simplification for semantic role labeling. In *Proceedings of the 46th Annual Meeting of the Association for Computational*

Linguistics (ACL) and the Human Language Technology Conference (HLT), pages 344–352.

vor der Brück, T., Hartrumpf, S., and Helbig, H. (2008). A readability checker with supervised learning using deep syntactic and semantic indicators. *Informatica*, 32(4):429–435.

Vossen, P. (1998). *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers.

Štajner, S., Drndarević, B., and Saggion, H. (2013). Corpus-based Sentence Deletion and Split Decisions for Spanish Text Simplification. *Computacion y Systemas*, 17(2):251–262.

Štajner, S., Evans, R., and Dornescu, I. (2014a). Assessing Conformance of Manually Simplified Corpora with User Requirements: the Case of Autistic Readers. In *Proceedings of the Workshop on Automatic Text Simplification – Methods and Applications in the Multilingual Society (ATS-MA)*, pages 53–63, Dublin, Ireland. ACL and Dublin City University.

Štajner, S., Evans, R., Orasan, C., and Mitkov, R. (2012). What Can Readability Measures Really Tell Us About Text Complexity? In *Proceedings of the LREC Workshop on Natural Language Processing for Improving Textual Accessibility (NLP4ITA)*, pages 14–21, Istanbul, Turkey.

Štajner, S., Mitkov, R., and Saggion, H. (2014b). One step closer to automatic evaluation of text simplification systems. In *Proceedings of the 3rd Workshop on Predicting*

and Improving Text Readability for Target Reader Populations (PITR), pages 1–10, Gothenburg, Sweden. ACL.

Štajner, S. and Saggion, H. (2013). Adapting Text Simplification Decisions to Different Text Genres and Target Users. *Procesamiento del Lenguaje Natural*, 51:135–142.

Vu, T. T., Tran, G. B., and Pham, S. B. (2014). Learning to simplify children stories with limited data. In *Intelligent Information and Database Systems (Part I)*, LNAI 8397, pages 31–41. Springer International Publishing Switzerland.

W3C (2008). *Web Content Accessibility Guidelines (WCAG) 2.0.*

Woodsend, K. and Lapata, M. (2011a). Learning to Simplify Sentences with Quasi-Synchronous Grammar and Integer Programming. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 409–420. ACL.

Woodsend, K. and Lapata, M. (2011b). WikiSimple: Automatic Simplification of Wikipedia Articles. In *Proceedings of the 25th AAAI Conference on Atificial Intelligence*, pages 929–932.

Wubben, S., van den Bosch, A., and Krahmer, E. (2012). Sentence simplification by monolingual machine translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers*, volume 1, pages 1015–1024, Stroudsburg, USA. ACL.

Yamada, K. and Knight, K. (2001). A syntax-based statistical translation model. In *Pro-*

ceedings of the 39th Annual Meeting on Association for Computational Linguistics (ACL), pages 523–530. ACL.

Yatskar, M., Pang, B., Danescu-Niculescu-Mizil, C., and Lee, L. (2010). For the sake of simplicity: unsupervised extraction of lexical simplifications from Wikipedia. In *Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 365–368. ACL.

Zhu, Z., Berndard, D., and Gurevych, I. (2010). A Monolingual Tree-based Translation Model for Sentence Simplification. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING)*, pages 1353–1361.

APPENDIX A

RELATED PUBLICATIONS

Some of the work described in this thesis is based on the previously published book chapter and articles in international journals or proceedings of peer-reviewed international conferences. Most of this work has been extended before its inclusion in this thesis:

1. Štajner, S., Evans, R., Orasan, C. and Mitkov, R. 2012. What can readability measures really tell us about text complexity? In *Proceedings of the Workshop on Natural Language Processing for Improving Textual Accessibility (NLP4ITA), held in conjunction with LREC 2012*. Istanbul, Turkey, May 27, pp. 14–21.

2. Drndarevic, B., Štajner, S. and Saggion, H. 2012. Reporting Simply: A Lexical Simplification Strategy for Enhancing Text Accessibility. In *Proceedings of the Easy-to-read on the Web Symposium*.

3. Drndarevic, B., Štajner, S., Bott, S., Bautista, S. and Saggion, H. 2013. Automatic Text Simplification in Spanish: A Comparative Evaluation of Complementing Modules. In *Proceedings of the 14th International Conference on Intelligent Text Processing and Computational Linguistics (Part II)*. Lecture Notes in Computer Science, Vol. 7817, Springer, pp. 488–500.

4. Štajner, S., Drndarevic, B. and Saggion, H. 2013. Corpus-based Sentence Deletion and Split Decisions for Spanish Text Simplification. *Computación y Sistemas*, Vol.17, No.2, pp. 251–262, ISSN 1405-5546.

5. Štajner, S. and Saggion, H. 2013. Readability Indices for Automatic Evaluation of Text Simplification Systems: A Feasibility Study for Spanish. To appear in *Proceedings of the 6th International Joint Conference on Natural Language Processing (IJCNLP 2013)*, Nagoya, Japan, 14–18 October 2013, pp. 374–382.

6. Štajner, S. and Saggion, H. 2013. Adapting Text Simplification Decisions to Different Text Genres and Target Users. *Procesamiento del Lenguaje Natural*, Vol. 51, pp. 135–142.

7. Glavaš G. and Štajner S. 2013. Event-Centered Simplification of News Stories. In *Proceedings of the Student Research Workshop at the International Conference on Recent Advances in Natural Language Processing (RANLP 2013)*, Hissar, Bulgaria, 9–11 September 2013, pp. 71–78. (**Best paper award**)

8. Štajner, S., Mitkov, R. and Saggion, H. 2014. One Step Closer to Automatic Evaluation of Text Simplification Systems. In Proceedings of the EACL workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR), Gothenburg, Sweden, 27 April 2014, pp. 1-10.

9. Mitkov, R. and Štajner, S. 2014. The Fewer, the Better? A Contrastive Study about Ways to Simplify. Proceedings of the COLING workshop on Automatic

Text Simplification - Methods and Applications in the Multilingual Society (ATS-MA), Dublin, Ireland, 24 August 2014, pp. 30-40.

10. Štajner, S., Evans R. and Dornescu, I. 2014. Assessing Conformance of Manually Simplified Corpora with User Requirements: the Case of Autistic Readers. Proceedings of the COLING workshop on Automatic Text Simplification - Methods and Applications in the Multilingual Society (ATS-MA), Dublin, Ireland, 24 August 2014, pp. 53-63.

11. Štajner, S. 2014. Translating sentences from 'original' to 'simplified' Spanish. Procesamiento del Lenguaje Natural, Vol. 53, pp. 61-68.

12. Štajner, S., Mitkov, R. and Corpas Pastor, G. 2014. 'Simple or not simple? A readability question'. In N. Gala, R. Rapp, and G. Bel-Enguix (eds), Recent Advances in Language Production, Cognition and the Lexicon, Springer, pp. 379-398.